DOCUMENT RESUME

ED 151 417 TM 006 951

AUTHOR Wilcox, Rand R.
TITLE Statistical Techniques for Criterion-Referenced Tests. Final Report. October, 1976-October, 1977.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C. Basic Skills Group.
PUB DATE 77
GRANT NIE-G-76-0083
NOTE 135p.

EDRS PRICE MF-$0.83 HC-$7.35 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; Cutting Scores; *Error Patterns; Item Sampling; *Mastery Tests; *Mathematical Models; *Statistical Analysis; *Test Interpretation; True Scores
IDENTIFIERS Bayesian Inference; Binomial Error Model; *Estimation; Monte Carlo Methods

ABSTRACT

Three statistical problems related to criterion-referenced testing are investigated: estimation of the likelihood of a false-positive or false-negative decision with a mastery test, estimation of true scores in the Compound Binomial Error Model, and comparison of the examinees to a control. Two methods for estimating the likelihood of false-positive and false-negative errors were developed and compared. For both procedures, it was found that when observations are generated according to a two-term approximation to the compound binomial, there was negligible effect on the accuracy of the estimates. Neither of the two estimation procedures consistently gave more accurate results. The estimation of true scores in the compound binomial error model was also explored. Findings showed that it was often possible to obtain a nontrivial improvement over the sample when simultaneously estimating the true score of several examinees; these estimation techniques appeared to be robust in terms of departure from the binomial distribution. The final investigation dealt with the standards used when making a mastery/nonmastery decision in terms of a particular skill. It was suggested that the true scores of the examinees be compared to a control population when defining minimum competency. (Author/MV)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Center for the Study of Evaluation

UCLA Graduate School of Education
Los Angeles, California 90024

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED151417

FINAL REPORT
OCTOBER, 1976 - OCTOBER, 1977

STATISTICAL TECHNIQUES FOR CRITERION-REFERENCED TESTS

(Grant # NIE-G-76-0083)

Rand R. Wilcox

TM006 951

Center for the Study of Evaluation

The Center for the Study of Evaluation (CSE) is a university-based research and development center established in 1966 by the Department of Health, Education, and Welfare. The mission of the Center has been to conduct research, development, training, and dissemination activities for the purpose of improving the planning and evaluation of education and other social action programs. More specifically, CSE activities include: guidance for educators and researchers in the areas of tests and measures and in educational decision making; development of easy to use evaluation procedures and materials in the form of manuals, workshops and kits; and evaluation of projects for federal, state, local, and university agencies requesting technical assistance in evaluation. The work of CSE during its first decade has clarified the ways in which systematic evaluation can be helpful in the improvement of many kinds of programs.

Information about CSE and its publications may be obtained by writing to:

Director, Public Information
Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California 90024

STATISTICAL TECHNIQUES FOR CRITERION-REFERENCED TESTS

This is the final report for the grant on Statistical techniques for criterion-referenced tests awarded to Dr. Wilcox of the Center for the Study of Evaluation by the Basic Skills Group of the National Institute of Education.

The grant supports the investigation of three statistical problems relating to criterion-referenced testing.

Problem #1. Estimating the likelihood of a false-positive or false-negative decision with a mastery test.

A false-positive error occurs with a criterion-referenced test when the examiner estimates an examinee's percent correct true score to be above a criterion-level π_o when in fact it is not. A false-negative error occurs when the examiner concludes that the examinee's percent correct true score is below π_o when the reverse is true. Two methods were developed and compared for estimating the likelihood of these two types of errors. Of particular interest was the effect of departures from the binomial distribution. For both estimation procedures it was found that when observations are generated according to a two-term approximation to the compound binomial, there was negligible effect on the accuracy of the estimates. It was also found that neither of the two estimation procedures dominated the other, i.e., always gave more accurate results. Conditions under which accurate estimates are obtained are described in Attachment A. This paper (Attachment A) has

been reviewed by Chester Harris, and Michael Subkoviak and it is scheduled to appear in the Winter 1977 issue of the Journal of Educational Statistics.

A paper describing a simplified method of how to estimate the likelihood of a false-positive or false-negative decision is described in separate report which is coauthored with Michael Subkoviak. (See Attachment B.) This paper has been submitted to the Journal of Educational Measurement for publication.

Problem #2. Estimating True Score in the Compound Binomial Error Model.

The second investigation deals with the problem of estimating true score in the compound binomial error model. Our major finding in this investigation is that it is often possible to obtain a nontrivial improvement over the sample when simultaneously estimating the true score of several examinees. One of these estimation techniques always performed about as well as the sample mean and so there appears to be little risk in using it in practice. Another important finding is that these estimation techniques appear to be robust in terms of departure from the binomial distribution. That is, if an estimator is developed assuming that the binomial error model holds, it will still give good results when observations are generated according to a two-term approximation to the compound binomial distribution. A paper describing our results has been reviewed by Chester Harris (see Attachment C) and is now under revision for publication in Psychometrika.

Problem #3. Comparing Examinees to a Control.

The final investigation deals with the problem of setting standards when making a mastery/nonmastery decision in terms of a particular skill.

More specifically, it was suggested that examinees be compared to a control population when defining minimum competency. As indicated in the original proposal this problem is intimately connected to the problem of specifying the length of a mastery test. When comparing examinees to a control it is desired to determine whether an examinee's percent correct true score is above or below the true score of the control. By choosing n, the test length, sufficiently large it is possible to guarantee that the examinees being compared to the control will be correctly classified with some pre-assigned probability level. The goal is to choose the smallest n so that a correct decision is made with probability at least P^*.

Using ranking and selection techniques, a general framework has been described for solving this problem. An expression has been derived which gives a lower bound on the probability of correctly classifying the examinees for a given n. An exact solution is now possible for most cases of interest including the case of normal populations having unknown and unequal variances. The one case we considered which does not have an exact solution is when the number of sampled control examinees is small and when the distribution of observed scores is not normal. In this case we have an approximate solution which appears to give conservative results. Attachment D, which has been reviewed by Chester Harris, describes our results in more detail. This paper has been reviewed and revised for publication in Psychometrika. Attachment E, which was reviewed by Milton Sobel, describes in more detail the solution for the case of normal populations. An abstract of this paper will appear in the Bulletin of the Institute of Mathematical Statistics.

ESTIMATING THE LIKELIHOOD OF FALSE-POSITIVE AND FALSE-NEGATIVE DECISIONS IN MASTERY TESTING: AN EMPIRICAL BAYES APPROACH

Rand R. Wilcox

Center for the Study of Evaluation
University of California, Los Angeles

ABSTRACT

False-positive and false-negative decisions are the fundamental errors committed with a mastery test; yet the estimation of the likelihood of committing these errors has not been investigated. Accordingly, two methods of estimating the likelihood of committing these errors are described and then investigated using Monte Carlo techniques. Conditions for obtaining accurate estimates are noted.

1. INTRODUCTION

Typically a mastery test is designed to sort k examinees into one of two mutually exclusive groups. For example, in a program of Individually Prescribed Instruction a student's progress through each level of a program of study is governed by his performance on a test dealing with individual behavioral objectives. The purpose of a test in such situations is to make a mastery/nonmastery decision for each of k examinees. If a mastery decision is made for a particular examinee, then he is advanced to the next level of instruction. If, however, a nonmastery decision is made, he will be given remedial work.

A model of mastery testing which is frequently adopted may be described as follows: A pool or domain of dichotomously scored test items, having mixed item difficulty, is constructed in relation to a particular course of instruction. The item pool may exist de facto or it may be a convenient conceptualization. The item form notion of Hively and others (1973) represents such a conceptualization. Let λ_i denote the percent correct domain or "true" score of the i^{th} examinee; λ_i represents the percent of items that the i^{th} examinee would answer correctly if he were to respond to every item in the item pool at a given occasion in time. With respect to the domain of items an examinee is said to have attained mastery if $\lambda_i \geq \lambda_0$ and nonmastery if $\lambda_i < \lambda_0$ where λ_0 is a known constant with a value between zero and one. The problem is to make a mastery/nonmastery decision for a given examinee based on his responses to n items randomly selected from the item domain. A mastery decision is made if an examinee answers n_0 or more items correctly where $0 \leq n_0 \leq n$. Note that λ_0 corresponds to a concept of mastery and n_0 to a mastery score or index (Harris, 1974). We further observe that the model of mastery testing just described is equivalent to the ranking and selection problem of partitioning k populations (examinees) with respect to a standard.

This model provides a reasonable description of mastery testing and is consistent with definitions of mastery or criterion-referenced tests (Glaser and Nitko, 1971; Harris, 1974). (See also Hambleton and Novick, 1973; Fhaner, 1974; Novick and Lewis, 1974; Huynh, 1976; Wilcox, 1976).

A false-positive error occurs when the examiner estimates an examinee's true score λ_i to be above the criterion level λ_0 when in fact it is not. A false-negative error occurs when λ_i is estimated to be below λ_0 when the reverse is true. False-positive and false-negative errors are the two errors that can be made in a two-valued classification, yet the estimation of the probability of committing these types of errors in connection with mastery testing has been virtually ignored. Instead attention has been given to measures of stability such as the proportion of agreement (Hambleton and Novick, 1973) which estimates the probability of randomly selecting an examinee and classifying him the same way based on two administrations of the same test. Certainly it is desirable to have a test with a high degree of stability. However, it may be that such a test is consistently inaccurate.

Let α and β equal the probability of committing a false-positive and false-negative decision, respectively, for an examinee chosen at random from some population of potential examinees. Observe that the values of both α and β are a function of the number of items on the test as well as the instructional history of the examinees. Consequently, knowing α and β provides a meaningful characterization of the entire teaching-testing complex. The purpose of this paper is to examine the problem of estimating α and β based on student response data.

In this paper, both α and β represent a joint probability. For example, α is the probability that a typical examinee will have $\lambda < \lambda_0$ and $x \geq n_0$ where x is the number correct observed score. In some situations a conditional probability may be more appropriate. Consider, for instance, the problem of determining the appropriate number of items to be administered to a single examinee. In the Bayesian solution given by Novick and Lewis (1974) a conditional probability is used, namely, the probability of making a correct decision given the examinee's observed score. Here, however, we are concerned with characterizing a specific group or population of examinees rather than one particular student and so defining α and β as joint probabilities is more convenient.

We observe that characterizing a typical or representative examinee is consistent with other measures devised to describe a mastery test. For example, the proportion of agreement (Hambleton and Novick, 1973) may be interpreted as an estimate of the probability of making a mastery-mastery or nonmastery-nonmastery decision for a randomly selected examinee based on two administrations of the same (or parallel) test. We hasten to add, however, that the proportion of agreement could be defined in terms of a single examinee. One might also follow Subkoviak (1976) who defines it in terms of a group of k examinees where k is a fixed constant.

Another instance where a mastery test is described in terms of a population of examinees is Harris' index of efficiency (Harris, 1974). This index estimates the squared point-biserial correlation coefficient between total test score and the mastery or nonmastery state of the examinees.

We also observe that α and β defined as joint probabilities play an important role in the paper by Huynh (1976) where the goal is to determine the optimal values of λ_0 and n_0. (See, in particular, Huynh's expression [8].) Within

the beta-binomial model adopted by Huynh, an estimate of both α and β is readily obtained as indicated in section 3 of this paper. What needs to be determined is the accuracy of this estimation procedure.

The binomial error model gives a reasonable approximation to the observed score distributions on tests (Lord, 1965, p. 253), but the compound binomial may be more realistic (Lord, 1965, Section 6; Lord and Novick, 1968, Chapter 23) and hence may give more accurate results. Accordingly, two methods of estimating α and β are described, and the accuracy of these statistics are examined under both the binomial and compound binomial error models. Sections 3 and 4 derive estimates of α and β assuming that the distribution of true scores belongs to a particular parametric family. Section 5 examines the accuracy of these estimation procedures using Monte Carlo techniques.

2. MATHEMATICAL STATEMENT OF THE PROBLEM

As indicated earlier, we let λ_i denote the proportion correct "true" score of an examinee taking an n item test. We regard λ_i $(i = 1,\ldots,k)$ as a sample from a prior distribution, say $g(\lambda)$, where $0 \leq \lambda \leq 1$. Let $h(x \mid \lambda_i)$ be the distribution of observed scores for a given true score λ_i. Assuming that $g(\lambda)$ is an integrable function and since $h(x \mid \lambda_i)$ is discrete, we have:

$$\alpha = \sum_{x=n_o}^{n} \int_{o}^{\lambda_o} h(x \mid \lambda)\, g(\lambda)\, d\lambda \tag{2.1}$$

and

$$\beta = \sum_{x=0}^{n_o-1} \int_{\lambda_o}^{1} h(x \mid \lambda)\, g(\lambda)\, d\lambda \tag{2.2}$$

If we knew $g(\lambda)$ and if we assume $h(x \mid \lambda)$ is binomial, we would also know α and β. However, $g(\lambda)$ is usually unknown. The approach taken here is to use empirical Bayes procedures to estimate $g(\lambda)$. When $h(x \mid \lambda)$ is assumed to be compound binomial, it too must be estimated. We are particularly interested in the accuracy of point estimates of α and β for relatively small values of k and n. We emphasize that the results given below do not reflect directly the accuracy of our estimate of $g(\lambda)$. In fact we are only concerned with an accurate estimate of $g(\lambda)$ in so far as it improves our estimate of α and β. It may be, for example, that a relatively poor estimate of $g(\lambda)$ will yield a reasonably accurate

estimate of the frequency of occurrence of both false-positive and false-negative decisions.

As a measure of the accuracy of any statistic $\hat{\alpha}$ which is used to estimate α we use

$$(2.3) \quad w_a^2 = E(\alpha - \hat{\alpha})^2$$

which represents the average or expected value of $(\alpha - \hat{\alpha})^2$ based on random samples of k examinees taking an n item test. Note that w_a is a function of the joint distribution of x and λ as well as k and n. Correspondingly, when estimating β, we use $w_b^2 = E(\beta - \hat{\beta})^2$.

3. ESTIMATION OF α AND β ASSUMING A BETA PRIOR

The estimation of the prior distribution $g(\lambda)$ is a most difficult problem for which no general solution exists. In this section we adopt a commonly used approach; namely, we assume λ has a beta distribution with parameters r and s. That is,

$$g(\lambda) = \frac{\Gamma(r+s)}{\Gamma(r)\,\Gamma(s)} \lambda^{r-1} (1-\lambda)^{s-1}$$

where Γ is the usual gamma function. This assumption is restrictive because there is little if any doubt that $g(\lambda)$ does not belong to the family of beta priors. However, there does not appear to be any compelling evidence that some other form of g should be used. Moreover there are several reasons for considering this case. First, the beta density is the natural conjugate prior of the binomial kernel $h(x \mid \lambda)$ (Raiffa and Schlaifer, 1961). Second, we have indentifiability, i.e., there exists a unique g such that

$$(3.1) \quad f(x) = \int_0^1 h(x \mid \lambda)\, g(\lambda)\, d\lambda$$

where f(x) is the marginal distribution of observed scores (Maritz, 1970, chapter 2). Third, there is evidence that a reasonable though not entirely satisfactory approximation of the true score distribution can be obtained with a two parameter beta prior (Keats and Lord, 1962). Finally, and perhaps most importantly of all, the results of estimation procedures assuming a particular parametric form for the prior may be used as a benchmark for judging alternate estimation techniques.

Let x_{ij} ($i = 1,\ldots,k$; $j = 1,\ldots n$) denote the j^{th} observation on the i^{th} examinee. By estimating the parameters r and s of the beta prior based on the sample x_{ij}, we obtain an estimate of g which in turn yields an estimate of both α and β. We begin by describing a method of estimating r and s which assumes that the binomial error model (Lord and Novick, 1968, chapter 23) holds, i.e.,

$$h(x \mid \lambda) = \binom{n}{x} \lambda^x (1-\lambda)^{n-x}$$

Let $M_{[t]}$ denote the t^{th} factorial moment of the marginal distribution of observed scores, i.e.,

$$M_{[t]} = \sum_{x=0}^{n} \frac{x!}{(x-t)!} f(x)$$

Let μ_t represent the t^{th} moment of the true score distribution $g(\lambda)$. Then

$$(3.2) \quad \mu_t = \frac{(n-t)!\, M_{[t]}}{n!}$$

(Lord and Novick, 1968, expression 23.8.4). We obtain unbiased estimates of $M_{[1]}$ and $M_{[2]}$ with

$$(3.3a) \quad \hat{M}_{[1]} = \frac{1}{k} \sum_{i=1}^{k} x_i$$

$$(3.3b) \quad \hat{M}_{[2]} = \frac{1}{k} \sum_{i=1}^{k} (x_i^2 - x_i)$$

where $x_i = \sum_{j=1}^{n} x_{ij}$. From (3.2) we have that

$$(3.4a) \quad \hat{\mu}_1 = \hat{M}_{[1]} / n$$

$$(3.4b) \quad \hat{\mu}_2 = \hat{M}_{[2]} / (n(n-1))$$

are unbiased estimates of μ_1 and μ_2. Thus, the mean, say μ and the variance, say σ^2, of the prior distribution may be estimated as

$$(3.5a) \quad \hat{\mu} = \hat{\mu}_1$$

(3.5b) $\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2$

Note that $\hat{\sigma}^2$ is a biased estimate of σ^2. Also note that it is possible to have $M_{[2]} = 0$ and $M_{[1]} > 0$ resulting in a negative estimate of σ^2. When this occurs we estimate each λ_i as

$$\hat{\lambda} = \Sigma_{i=1}^{k} \Sigma_{j=1}^{n} x_{ij}/(kn)$$

For this special condition, if $\hat{\lambda} \geq \lambda_o$ we estimate α to be zero and β to be $\sum_{x=0}^{n_o-1} \binom{n}{x} (\hat{\lambda})^x (1-\hat{\lambda})^{n-x}$. Correspondingly, if $\hat{\lambda} < \lambda_o$, we estimate β to be zero and α to be

$$\Sigma_{x=n_o}^{n} \binom{n}{x} (\hat{\lambda})^x (1-\hat{\lambda})^{n-x}.$$

In terms of r and s

(3.6a) $$\mu = \frac{r}{r+s}$$

(3.6b) $$\sigma^2 = \frac{rs}{(r+s)^2 (r+s+1)}$$

(Johnson and Kotz, 1970). Solving (3.6) for r and s gives

(3.7a) $$r = \frac{\mu^2 (1-\mu)}{\sigma^2} - \mu$$

(3.7b) $$s = \frac{\mu (1-\mu)^2}{\sigma^2} + \mu - 1$$

Substituting in (3.7) $\hat{\mu}$ and $\hat{\sigma}^2$ for μ and σ^2, respectively, yields an estimate of r and s, say $\hat{r}$ and $\hat{s}$. The estimates of r and s may then be used to estimate α and β as

(3.8a) $$\hat{\alpha}_1 = \sum_{x=n_o}^{n} \int_o^{\lambda_o} \binom{n}{x} \lambda^x (1-\lambda)^{n-x} \frac{\Gamma(\hat{r}+\hat{s})}{\Gamma(\hat{r})\,\Gamma(\hat{s})} \lambda^{\hat{r}-1} (1-\lambda)^{\hat{s}-1}\, d\lambda$$

$$(3.8b)\quad \hat{\beta}_1 = \sum_{x=0}^{n_o-1} \int_{\lambda_o}^{1} \binom{n}{x} \lambda^x (1-\lambda)^{n-x} \frac{\Gamma(\hat{r}+\hat{s})}{\Gamma(\hat{r})\,\Gamma(\hat{s})} \lambda^{\hat{r}-1}(1-\lambda)^{\hat{s}-1}\, d\lambda$$

A referee pointed out that maximum likelihood estimates of r and s can be obtained iteratively as indicated by Griffiths (1973). However, we do not consider this method of estimating r and s here. Due to the complexity of the statistics used to estimate α and β, Monte Carlo procedures were used to assess the accuracy of (3.8a) and (3.8b). The results are reported in section 5. The cost of these studies was immense. Unfortunately, our computer budget is finite and so we were forced to exclude from consideration all iterative estimation techniques.

As indicated earlier, the binomial error model may not be completely satisfactory in an item sampling model. As suggested by Lord and Novick (1968, chapter 23) we use a two-term approximation to the compound binomial, viz.,

$$(3.9)\quad \bar{h}(x \mid \lambda) = p_n(x) + d\lambda(1-\lambda)\, C(x)$$

where

$$p_n(x) = \binom{n}{x} \lambda^x (1-\lambda)^{n-x}, \quad x = 0,1,\ldots,n$$

$$= 0, \qquad \text{otherwise}$$

$$C(x) = \sum_{v=0}^{2} (-1)^{v+1} \binom{2}{v} p_{n-2}(x-v)$$

The parameter d is given by

$$(3.10)\quad d = \frac{n^2 (n-1)\, \sigma_\pi^2}{2[\mu_x (n-\mu_x) - \sigma_x^2 - n\,\sigma_\pi^2]}$$

where σ_x^2 and μ_x are the population variance and mean of the distribution of observed scores and where σ_π^2 is the population variance of the item difficulties. We observe that d is chosen so that the correlation between observed score and true score will always equal the square root of the Kuder-Richardson formula 20 reliability coefficient. Lord (1965) notes that (3.9) is a close approximation to a frequency distribution for most cases of interest. Difficulties could

arise if d were too large; we avoid these difficulties by assuming $0 \leq d \leq 4.0$. For all 16 distributions reported by Lord, the values of d were in this range. (See Lord, 1965, p. 264.)

Under the more general compound binomial we are still able to estimate α and β. As shown by Lord (1965, p. 265) the mean and variance of the distribution of true scores for the two-term approximation to the compound binomial error model are given by:

$$(3.11) \quad \mu = M_{[1]}/n$$

$$(3.12) \quad \sigma^2 = [\sigma_x^2 - (n-2d)\,\bar{p}\,\bar{q}\,] \,/\, [n(n-1) + 2d]$$

where $\bar{p} = M_{[1]}/n$ and $\bar{q} = 1-\bar{p}$.

d may be estimated using standard item analysis techniques. Hence, the parameters r and s of the beta prior may be estimated using (3.7) above.

Substituting (3.9) into (3.8a), the estimate of α under the compound binomial error model is

$$(3.13a) \quad \hat{\alpha}_{1(d)} = \sum_{x=n_o}^{n} \int_0^{\lambda_o} \bar{h}(x \mid \lambda)\, \frac{\Gamma(\hat{r}+\hat{s})}{\Gamma(\hat{r})\,\Gamma(\hat{s})}\, \lambda^{\hat{r}-1}(1-\lambda)^{\hat{s}-1}\, d\lambda$$

Correspondingly, we estimate β to be

$$(3.13b) \quad \hat{\beta}_{1(d)} = \sum_{x=0}^{n_o-1} \int_{\lambda_o}^{1} \bar{h}(x \mid \lambda)\, \frac{\Gamma(\hat{r}+\hat{s})}{\Gamma(\hat{r})\,\Gamma(\hat{s})}\, \lambda^{\hat{r}-1}(1-\lambda)^{\hat{s}-1}\, d\lambda$$

Before continuing, we should comment that Maritz (1970) suggests two alternative methods for estimating g which may be applied to the problem at hand. The first method uses a step function approximation of g. A polygon approximation is used in the other. Both of these procedures would seem to be appealing since they do not make any assumptions about the form of g. The difficulty is that both procedures require the maximization of a rather complicated function which requires an iterative solution. Originally we intended to include these procedures in our Monte Carlo experiments. It soon became apparent, however, that these iterative solutions increased our computer costs well beyond our means. (Note

that the Monte Carlo studies reported in Maritz's Table 3.10, p. 69 are not very extensive and they are based on only 20 iterations.)

4. ESTIMATION OF α AND β USING AN INVERSE SINE TRANSFORMATION

In the previous section a procedure for estimating α and β was described which is contingent upon estimating the parameters r and s of an assumed beta prior. One difficulty with this estimation procedure is that the statistics $\hat{r}$ and $\hat{s}$ no doubt lack the desirable properties of unbiasedness, maximum likelihood, and efficiency. Consequently, one might expect estimates of r and s to be poor for relatively small samples. We propose, therefore, to investigate the use of an inverse sine transform which converts a binomial random variable into an approximately normally distributed random variable with known variance, the variance being independent of the value of λ_i. This is often called a variance stabilizing transformation. Our main justification for considering the use of such a transformation is that it has been both useful and successful in similar situations. (See, e.g., Novick, Lewis and Jackson, 1973; Efron and Morris, 1975; Huynh, 1976.) The disadvantage of this approach is that the transformed random variable is asymptotically normal and thus any estimation procedure using small samples may be poor. In addition, the rate of convergence to normality is a function of the unknown parameter λ_i. The crucial question is, of course, whether this approach reduces the value of w_a and w_b as defined by (2.3).

Let

$$(4.1)\quad y_i = \tfrac{1}{2}(4n+2)^{\frac{1}{2}}\left(\sin^{-1}\left(\sqrt{\frac{x_i}{n+1}}\right) + \sin^{-1}\left(\sqrt{\frac{x_i+1}{n+1}}\right)\right)$$

where angles are in radians and, as before, $x_i = \sum_{j=1}^{n} x_{ij}$ is the observed score of the i^{th} examinee. The transformation (4.1) is suggested by Freeman and Tukey (1950) where y_i given λ_i is approximately normally distributed with mean

$$\tau = (4n+2)^{\frac{1}{2}} \sin^{-1}\left(\sqrt{\lambda_i}\right)$$

and variance one. As in the previous section we let μ and

σ^2 represent the mean and variance of the prior distribution where now the prior refers to the distribution of τ, not λ. As is customary (Raiffa and Schlaifer, 1961), we assume the prior has a normal distribution. Our concerns about the beta prior used in the previous section apply to the normal prior used here. The assumption of a normal prior implies that the marginal distribution of observed scores is also normally distributed with mean μ and variance $\sigma^2 + 1$. It follows that

$$E_T\, E(y \mid \tau) = E(\tau)$$

$$E_T\, E(y^2 - 1 \mid \tau) = E(\tau^2)$$

where E_T means expectation over the distribution of τ. Hence, we may estimate $\mu = E(\tau)$ and $\sigma^2 = E(\tau^2) - E^2(\tau)$ with

$$(4.2a)\quad \hat{\mu} = \frac{1}{k} \sum_{i=1}^{k} y_i$$

$$(4.2b)\quad \hat{\sigma}^2 = \frac{1}{k} \sum_{i=1}^{k} (y_i^2 - 1) - \hat{\mu}^2$$

It is known (see, for example, Hogg and Craig, 1970, p. 210) that the joint probability density function of y and λ is bivariate normal with common mean μ, respective variances $1 + \sigma^2$ and σ^2, and correlation $\sigma/\sqrt{1 + \sigma^2}$. Applying the method of moments, as was done in the previous section, the estimates of α and β are

$$(4.3a)\quad \hat{\alpha}_2 = \int_{\gamma}^{\infty} \int_{-\infty}^{\lambda'} \frac{1}{2\pi\hat{\sigma}} \exp\left\{\frac{-(y-\lambda)^2}{2} - \frac{(\lambda-\hat{\mu})^2}{2\hat{\sigma}^2}\right\} d\lambda\, dy$$

$$(4.3b)\quad \hat{\beta}_2 = \int_{-\infty}^{\gamma} \int_{\lambda'}^{\infty} \frac{1}{2\pi\hat{\sigma}} \exp\left\{\frac{-(y-\lambda)^2}{2} - \frac{(\lambda-\hat{\mu})^2}{2\hat{\sigma}^2}\right\} d\lambda\, dy$$

where

$$\gamma = \tfrac{1}{2}(4n + 2)^{\frac{1}{2}} \left[\sin^{-1}\left(\sqrt{\frac{n_o}{n+1}}\right) + \sin^{-1}\left(\sqrt{\frac{n_o + 1}{n+1}}\right)\right]$$

and

$$\lambda' = (4n + 2)^{\frac{1}{2}} \sin^{-1}\left(\sqrt{\lambda_o}\right)$$

//

Again we have the difficulty that $\hat{\sigma}^2$ may be negative.

In this case we set $\hat{\alpha}_2 = 0$, $\hat{\beta}_2 = \sum_{x=0}^{n_0-1} \binom{n}{x} \hat{\lambda}^x (1-\hat{\lambda})^{n-x}$ when

$\hat{\lambda} \geq \lambda_0$; and when $\hat{\lambda} < \lambda_0$ we set $\hat{\beta}_2 = 0$,

$\hat{\alpha}_2 = \sum_{x=n_0}^{n} \binom{n}{x} \hat{\lambda}^x (1-\hat{\lambda})^{n-x}$ as was done in the previous section.

It may be helpful to indicate how (4.3a) and (4.3b) can be evaluated with existing computer subroutines. For convenience we write the estimated bivariate distribution of y and λ as $\hat{h}(y \mid \lambda)\ \hat{g}(\lambda)$. Observe that

$$(4.4)\quad \int_{-\infty}^{\gamma} \int_{-\infty}^{\lambda'} \hat{h}(y \mid \lambda)\ \hat{g}(\lambda)\ d\lambda\ dy + \int_{-\infty}^{\gamma} \int_{\lambda'}^{\infty} \hat{h}(y \mid \lambda)\ \hat{g}(\lambda)\ d\lambda\ dy = \int_{-\infty}^{\gamma} \hat{f}(y)\ dy$$

where $\hat{f}(y)$ is the estimated normally distributed marginal distribution of observed scores y with mean $\hat{\mu}$ and variance $1 + \hat{\sigma}^2$. The first integral on the left hand side of (4.4) can be evaluated with the IMSL subroutine (1975) MDBNOR after the random variables y and λ are transformed so as to have common mean zero and variance one. The right hand side of (4.4) can be evaluated with the FORTRAN subroutine ERFC; thus, we have the value $\hat{\beta}_2$. The statistic $\hat{\alpha}_2$ may be evaluated in a similar manner.

The transformation (4.1) is claimed by Mosteller and Youtz (1961) as well as Mosteller and Tukey (1968) to be the best existing angular transformation for the binomial distribution. As indicated above, however, the compound binomial may be a more appropriate probability distribution for describing the observed frequency of test scores. To be conservative we introduce an inverse sine transformation for the compound binomial. The desirability of using this transformation will be discussed in section 5 below.

12

As shown by Lord (1965, p. 265) the mean and variance of the two term approximation to the compound binomial distribution given by (3.12) may be written as $n\lambda$ and $(n-2d)\ \lambda\ (1-\lambda)$. It follows that

$$(4.5)\quad \frac{2n}{\sqrt{n-2d}} \sin^{-1}\left(\sqrt{x/n}\right)$$

is asymptotically normal with mean $\dfrac{2n \sin^{-1}\left(\sqrt{\lambda}\right)}{\sqrt{n-2d}}$ and variance one (Rao, 1973, Section 6g). After estimating d via (3.11), one may use transformation (4.5) in place of (4.1), then estimate μ and σ^2 with (4.2a) and (4.2b), respectively, and finally estimate α and β with (4.3a) and (4.3b). When α and β are estimated using (4.5) for a given value of d, we denote the estimates by $\hat{\alpha}_{2(d)}$ and $\hat{\beta}_{2(d)}$, respectively.

5. METHOD AND RESULTS OF MONTE CARLO EXPERIMENTS

The marginal distribution of observed scores, f, can be computed using the results given by Lord (1965, p. 268). Since f is a step function, test scores corresponding to k randomly sampled examinees can be generated as follows: Generate a random number having a value between 0.0 and 1.0, i.e., a number from the uniform distribution. We accomplished this with IBM's SSP subprogram RANDU. Next, determine the x such that the generated random number is in the half-open interval $[f(x-1), f(x))$. This x is the observed test score of the first examinee. By repeating this process k times, we obtain observed scores which arise from a particular beta-compound binomial model.

Estimation of α and β using $\hat{\alpha}_1$, $\hat{\beta}_1$, $\hat{\alpha}_2$ and $\hat{\beta}_2$ based on these k generated observed scores constituted one iteration in a particular Monte Carlo experiment. For each Monte Carlo experiment performed, 1000 iterations were used. The 1000 resulting estimates of α and β using a particular estimation procedure were then used to estimate w_a and w_b.

In an earlier version of this paper we used only 500 iterations and a different method for generating observed scores. However, the Monte Carlo procedure described above is much more satisfactory for present purposes.

The values of r and s were chosen so as to include a variety of priors, namely, J-shaped, U-shaped, symmetric and

skewed distributions. Some of these distributions (e.g., U-shaped) are probably unrealistic in terms of mastery testing. They were included, however, so as to obtain more general results. Over 200 Monte Carlo studies were made using various values of r, s, d and λ_o.

Initially we set $\lambda_o = 0.7$ and $n_o = \lambda_o n$. For each prior distribution used, w_a and w_b were estimated by first using the exact value of d and then by setting d arbitrarily equal to zero. The values of k and n were (k, n) = (10, 10), (10, 20), (10, 30), (20, 10), (30, 10). For simplicity we discuss the results in terms of w_a. No additional insights were found when examining w_b.

Regardless of the true score distribution used, the value of d had negligible effect on the value of w_a when (3.8a) was used to estimate α, particularly for n=30. This result is illustrated in Table I for the special case (r, s) = (9, 2), $\lambda_o = .7$.

TABLE I

Values of w_a for r=9, s=2, λ_o=.7

d\k,n	10, 10	10, 20	10, 30	20, 10	30, 10
			Results using (3.8a)		
0.0	.082	.053	.038	.071	.069
2.5	.095	.056	.041	.099	.103
4.0	.098	.059	.042	.106	.107
			Results using (3.13a)		
2.5	.065	.043	.039	.069	.064
4.0	.048	.041	.039	.066	.064
			Results using (4.3a)		
0.0	.066	.033	.026	.039	.038
2.5	.061	.034	.026	.049	.047
4.0	.058	.035	.027	.055	.055
			Results using (4.5)		
2.5	.027	.024	.023	.021	.017
4.0	.031	.024	.023	.028	.027

Table I also illustrates that using the exact value of d in (3.13a) tends to have little effect on lowering w_a. As might be expected, the most noticeable improvement occurred for n=10.

As for $\hat{\alpha}_2$, again it was found that altering d had little effect on the value of w_a. Using the exact value of d in (4.5) tended to give better results than (4.1) for n=10. However, the improvement of (4.5) over (4.1) appears to decrease rapidly as n gets large.

Table I suggests that reasonably accurate estimates of α can be obtained particularly if n is greater than or equal to 30. To better assess the accuracy of $\hat{\alpha}_1$ and $\hat{\alpha}_2$ we present Table II which gives the values of w_a for both $\hat{\alpha}_1$ and $\hat{\alpha}_2$ where r=9, s=2, λ_o=.5(.1).8. From Table II we see that $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are very accurate for λ_o=.5 but that this accuracy diminishes considerably as λ_o approaches .8. Apparently the accuracy of both $\hat{\alpha}_1$ and $\hat{\alpha}_2$ depends on the actual value of α. Our findings suggest that as α gets large, w_a will increase for fixed k and n. In Table III, for example, we see that α=.005 for k=n=10, λ_o=.5. For λ_o=.8, α=.157. From Table II we see that the corresponding values of w_a are .055 and .130.

TABLE II

Values of w_a, r=9, s=2, d=0

λ_o \ k,n	10, 10	10, 20	10, 30	20, 10	30, 10
		Results using $\hat{\alpha}_1$			
.5	.055	.043	.033	.048	.037
.6	.085	.059	.043	.074	.059
.7	.082	.053	.038	.071	.069
.8	.130	.071	.047	.081	.072
		Results using $\hat{\alpha}_2$			
.5	.008	.006	.005	.006	.005
.6	.021	.015	.013	.021	.024
.7	.066	.033	.026	.039	.038
.8	.123	.075	.047	.070	.054

TABLE III

Actual Values of α

λ_o \ k,n	10, 10	10, 20	10, 30	20, 10	30, 10
.5	.005	.004	.004	.005	.005
.6	.022	.018	.015	.022	.022
.7	.067	.052	.045	.067	.067
.8	.157	.116	.096	.157	.157

Note that for λ_o=.8 the most effective method of lowering w_a is to increase n (the number of items) as opposed to increasing k (the number of individuals). In general, but not always, increasing n will decrease the value of α. Consequently, we lower the value of w_a by increasing n primarily because we obtain more accurate estimates of α when α is small. Increasing k with n fixed also lowered w_a but at a somewhat slower rate.

We also observed that neither of the statistics $\hat{\alpha}_1$ or $\hat{\alpha}_2$ dominated the other, i.e., had consistently lower values for w_a. Consequently, based purely on statistical considerations, it is impossible to recommend a particular estimation procedure. However $\hat{\alpha}_2$ usually performed at least as well as $\hat{\alpha}_1$, and frequently it gave substantially better results.

An Illustration

To illustrate the use and interpretation of estimates of α and β, we reanalyze some test data reported by Huynh (1976, p. 275). A five-item arithmetic test was administered to a group of 91 students. Here we set n_o=4, λ_o=.8. The estimates of r and s were $\hat{r}$=1.61 and $\hat{s}$=.857. For the beta-binomial model (i.e., d=0), we have $\hat{\alpha}_1$ = .1799 and $\hat{\beta}_1$=.0310. The examiner must decide whether α=.1799 and β=.031 are both acceptable. If, for example, a value for α greater than .1 is intolerable, the test results suggest that the test must be modified in some way. One possibility is to increase n_o which will lower α but increase β. If we set n_o=5, our estimates of α and β are now $\hat{\alpha}_1$=.054 and $\hat{\beta}_1$=.127. Another possibility is to increase the test length which will lower both α and β. If we use n=10, n_o=8, we get $\hat{\alpha}_1$=.1167 and $\hat{\beta}_1$=.0320. For n=20, n_o=16, we have $\hat{\alpha}_1$=.075 and $\hat{\beta}$=.029.

ACKNOWLEDGEMENTS

The work upon which this publication is based was performed pursuant to Grant # NIE-G-76-0083 with the National Institute of Education, Department of Health, Education and Welfare. Points of view or opinions stated do not necessarily represent official NIE position or policy. The author would like to thank both referees, the associate editor, Chester Harris and Michael Subkoviak for their most helpful comments on earlier drafts of this paper.

REFERENCES

Efron, B., & Morris, C. Data analysis using Stein's estimation and its generalizations. Journal of the American Statistical Association, 1975, 70, 311-319.

Fhaner, S. Item sampling and decision making in achievement testing. British Journal of Mathematical and Statistical Psychology, 1974, 27, 172-175.

Freeman, M. F., & Tukey, J. W. Transformations related to the angular and the square root. The Annals of Mathematical Statistics, 1950, 21, 607-611.

Glaser, R., & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington: American Council on Education, 1971.

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Harris, C. W. Some technical characteristics of mastery tests. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced measurement. CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Hively, W., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S. Domain-referenced curriculum evaluation: A technical handbook and a case study from the MINNEMAST project. CSE Monograph Series in Evaluation, No. 1. Los Angeles: Center for the Study of Evaluation, University of California, 1973.

Hogg, R. V., & Craig, A. T. Introduction to mathematical statistics. New York: Macmillan, 1970.

Huynh, H. On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.

Huynh, H. Statistical consideration of mastery scores. Psychometrika, 1976, 41, 65-78.

IBM Application Program, System/360. Scientific subroutines package (360-CM-03X) Version III, programmer's manual. White Plains, New York: IBM Corporation Technical Publications Department, 1971.

IMSL Library 1, Vol. II. Houston: International Mathematical and Statistical Libraries, 1975.

Keats, J. A., & Lord, F. M. A theoretical distribution for mental test scores. Psychometrika, 1962, 27, 59-72.

Lord, F. M. A strong true-score theory, with applications. Psychometrika, 1965, 30, 239-270.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass: Addison-Wesley, 1968..

Maritz, J. S. Empirical Bayes methods. London: Methuen, 1970.

Mosteller, F., & Tukey, J. W. Data analysis, including statistics. In G. Lindzey & E. Aronsen (Eds.), The Handbook of Social Psychology. Reading Mass: Addison-Wesley, 1968.

Mosteller, F., & Youtz, C. Tables of the Freeman-Tukey transformations for the binomial and Poisson distributions. Biometrika, 1961, 48, 433-440.

Novick, M. R., & Lewis, C. Prescribing test length for criterion-referenced measurement. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced measurement. CSE Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Novick, M. R., Lewis, C., & Jackson, P. E. The estimation of proportions in m groups. Psychometrika, 1973, 38, 19-46.

Raiffa, H., & Schlaifer, R. Applied statistical decision theory. Boston: Division of Research, Graduate School of Business Administration, Harvard University, 1961.

Rao, C. R. Linear statistical inference and its applications. New York: John Wiley, 1973.

Subkoviak, M. J. Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 1976, 13, 265-276.

Wilcox, R. R. A note on the length and passing score of a mastery test. Journal of Educational Statistics, 1976, 1, 359-364.

AUTHOR

WILCOX, RAND R. Address: Center for the Study of Evaluation, UCLA Graduate School of Education, 405 Hilgard Avenue, Los Angeles, CA 90024. Title: Senior Research Associate. Degrees: B.A., M.A., Ph.D., University of California. Specialization: Educational measurement and statistical methods.

Estimating the Probability of Correct Classification in Mastery Testing

Michael J. Subkoviak
University of Wisconsin

Rand Wilcox
University of California at Los Angeles

Running head: Probability of Correct Classification in Mastery Testing

Abstract

A procedure is proposed for estimating the proportion of persons in a group that are correctly classified on a mastery test, i.e., the proportion whose observed classification agrees with their true classification. A numerical example is provided, and extensions of the procedure are discussed.

Estimating the Probability of Correct Classification in Mastery Testing

Keats and Lord (1962) have proposed a relatively simple mathematical model for test scores that has a number of practical applications. For instance, Huynh (1976) and Subkoviak and Albrecht (1977) have demonstrated empirically that the model can be used to estimate the degree to which persons are consistently classified as masters or nonmasters on parallel mastery tests, where a pass-fail test with a cut-off score of 75% correct is one example of a mastery test.

The purpose of the present paper is to illustrate that the Keats-Lord model is also useful for estimating the extent to which persons are correctly classified, i.e., for estimating the proportion of persons whose classification based on observed score agrees with their classification based on true score. In addition, extensions of the procedure to the case of polychotomous classification and further generalizations of the Keats-Lord model are noted.

The Keats-Lord Model

Let us begin with the following notational definitions:

n = number of test items;

x = number of items correctly answered by an individual;

p = unknown true proportion correct score for an individual, i.e., the mean of an individual's observed proportion correct scores (x/n) over repeated parallel tests;

π = mastery criterion expressed as a proportion, e.g., $p \geq \pi$ implies true mastery and $p < \pi$ implies true nonmastery;

c = mastery criterion expressed as the number of items correct on an n-item test, e.g., $x \geq c$ implies observed mastery and $x < c$ implies observed nonmastery; c equals the smallest integer greater than or equal to $n\pi$.

Of interest here is the extent to which persons' observed mastery-nonmastery classifications (i.e., $x \geq c$ or $x < c$) are correct or, in other words, agree with their true mastery-nonmastery states (i.e., $p \geq \pi$ or $p < \pi$). One natural index of agreement in this sense is the probability that the observed classification corresponds to the true classification for a typical examinee. This probability will be symbolized Q, and thus Q represents the probability of correct classification for an examinee randomly selected from the population of potential examinees. Mathematically Q can be expressed as:

$$Q = P(x < c, p < \pi) + P(x \geq c, p \geq \pi), \qquad (1)$$

where $P(x < c, p < \pi)$ and $P(x \geq c, p \geq \pi)$ are respectively the probability of correct nonmastery and correct mastery classifications.

Under the assumptions of the Keats-Lord model (discussed below), it can be shown (see the Appendix) that Q in Equation 1, the probability of a correct classification, is given by:

$$Q = \{1/B(\ell+1, m+1)\}\{\sum_{x=0}^{c-1}\binom{n}{x} B(\ell+x+1, m+n-x+1)\, I_\pi(\ell+x+1, m+n-x+1) + \sum_{x=c}^{n}\binom{n}{x} B(\ell+x+1, m+n-x+1)[1-I_\pi(\ell+x+1, m+n-x+1)]\} \qquad (2)$$

If the mean (u), variance (σ^2), and Kuder-Richardson 21 reliability coefficient (ρ) of observed x-scores are estimated from a reasonably large sample of testees, the various terms in Equation 2 can be evaluated as follows:

n = number of test items;

π = mastery criterion expressed as a proportion;

c = smallest integer greater than or equal to $n\pi$;

$\rho = [n/(n-1)][1-u(n-u)/(n\sigma^2)]$;

$\ell = u(1/\rho-1)-1$;

$m = (n-u)(1/\rho-1)-1$;

$\binom{n}{x} = n!/[(n-x)!x!]$;

$B(\ ,\ )$ = beta function which can be calculated by common computer routines or can be found in standard mathematical tables;

$I_\pi(\ ,\ )$ = incomplete beta function which can also be computed by common computer routines or can be found in standard mathematical tables.[1]

The Keats-Lord model upon which Equation 2 is based involves two basic assumptions about the nature of true and error scores. It should be noted in passing that the model and its assumptions have been shown to fit a variety of real data sets quite adequately (Keats & Lord, 1962). First, it is assumed that the distribution of true scores (p) for the population of examinees is some member of the beta family of distributions. This family includes distributions having the usual bell-shape, as well as rectangular and U-shapes (see LaValle, 1970, p. 256 for more examples). As such, the beta assumption is quite liberal in that it accomodates a wide range of possible true score distributions.

Second, if n-item tests were repeatedly administered to a single individual with true score p, it is assumed that his or her distribution of observed scores (x) would be binomial with parameters n and p. This assumption tends to follow, for instance, if items are scored 0 and 1; if the outcome on one item does not affect the outcome on others; and if items are equally difficult. While the latter two conditions rarely, if ever, occur in practice, the quality of empirical results reported by Keats and Lord (1962) seems to indicate that the model is robust with respect to such violations.

Example

Suppose an n=4 item test was administered to N=100 students; and suppose the test scores $x = 0,1,2,3$ and 4 occurred with frequencies $f(x) = 8,25,35,25$ and 8 respectively. The sample mean and variance are $\hat{u} = \Sigma x f(x)/N = 2.00$ and $\hat{\sigma}^2 = \Sigma(x-\hat{u})^2 f(x)/(N-1) = 1.15$; and the estimate of KR21 is $\hat{p} = [n/(n-1)][1-\hat{u}(n-\hat{u})/(n\hat{\sigma}^2)] = .17$. The quantities ℓ and m required by Equation 2 are equal in this particular instance, i.e., $\hat{\ell} = \hat{u}(1/\hat{\rho}-1)-1 = 8.76$ and $\hat{m} = (n-\hat{u})(1/\hat{\rho}-1)-1 = 8.76$; but this is not true in general. Now if the mastery criterion is set at, say, $\pi = .85$, then $c = 4$ in Equation 2; since 4 is the smallest integer greater than or equal to $n\pi = 3.40$.

Summarizing, the parameters required by Equation 2 are: n=4, $\hat{\ell}=\hat{m}$=8.76, π=.85 and c=4. Thus by Equation 2 the probability of a correct classification is:

$$Q = \{1/B(9.76,\ 9.76)\}\{\sum_{x=0}^{3}\binom{4}{x}B(9.76+x,\ 13.76-x)I_{.85}(9.76+x,\ 13.76-x) +$$

$$\sum_{x=4}^{4}\binom{4}{x}B(9.76+x,\ 13.76-x)[1-I_{.85}(9.76+x,\ 13.76-x)]\}$$

where $\binom{4}{x}$, $B(\ ,\ )$ and $I_{.85}(\ ,\ )$ can be obtained via computer or standard mathematical tables (e.g., Pearson, 1956) and are as shown in Table 1. Substituting

Insert Table 1 about here

these values into the above equation gives the following result:

$$Q = \{1/(.15\times10^{-5})\}\{(1)(.12\times10^{-6})(.99)+(4)(.95\times10^{-7})(.99)+(6)(.87\times10^{-7})(.99)+$$
$$(4)(.95\times10^{-7})(.99)+(1)(.12\times10^{-6})(1-.99)\}$$

$$= .93.$$

Thus, it is likely that 93 out of the 100 students tested are correctly classified.

Discussion

In the example above, Q=.93 when the mastery criterion is set at π=.85. However, larger or smaller values of Q could be obtained for this same data set by simply changing π. In fact, the magnitude of Q is affected by a number of such factors. Specifically, the probability of correct classification, Q, tends to increase: (a) as the density of true scores about the mastery criterion decreases and (b) as the number of test items increases. Thus, Q would assume larger values for settings of the π-criterion in the tails of a bell-shaped distribution (as in the example above) than for settings in the middle of the distribution. Secondly, for a particular setting of π such as .85, Q would be larger for a $2n$-item test than for an n-item test.

In regard to generalization, it is worth noting that the probability of correct classification can also be estimated for the case of three or more levels of mastery. For example, let π_1 and π_2 (where $\pi_1 < \pi_2$) represent different degrees of mastery on the true score scale; and let c_1 and c_2 be the corresponding criteria on the observed score scale. Three categories of mastery are thus possible, and Equation 1 can be generalized as follows:

$$Q = P(x<c_1,\ p<\pi_1) + P(c_1 \leq x < c_2,\ \pi_1 \leq p < \pi_2) + P(c_2 \leq x,\ \pi_2 \leq p). \qquad (3)$$

The argument outlined in the Appendix can then be applied to Equation 3 to obtain an expression similar to Equation 2.

Extensions of the Keats-Lord model, on which Equation 2 is based, are also possible. For example, Lord (1965) has suggested a slightly more complex model that involves somewhat more general and reasonable assumptions regarding true and error score distributions (see also Lord, 1969). However, a Monte Carlo study by Wilcox (1977) seems to suggest that the more complex model adds little in the way of accuracy to probability estimates, which are of interest here. Of course this does not imply that the more complex model lacks utility for

certain other purposes, e.g., for simulating or estimating complete bivariate distributions of true and observed scores.

Appendix

The beta assumption of the Keats-Lord model implies that a true score, $\underline{p}$, can assume any value within the interval 0 to 1; while the binomial assumption implies that an observed score, $\underline{x}$, takes on only integral values between 0 and $\underline{n}$. Thus, Equation 1 can be written as a double summation of $\underline{P}(\underline{x},\underline{p})$ values, where $\underline{P}(\underline{x},\underline{p})$ represents the joint probability distribution of examinees' observed and true scores:

$$\underline{Q} = \sum_{\underline{x}=0}^{\underline{c}-1} [\int_0^{\pi} \underline{P}(\underline{x},\underline{p})d\underline{p}] + \sum_{\underline{x}=\underline{c}}^{\underline{n}} [\int_{\pi}^{1} \underline{P}(\underline{x},\underline{p})d\underline{p}], \tag{4}$$

where $\underline{P}(\underline{x},\underline{p}) = [\binom{\underline{n}}{\underline{x}} / \underline{B}(\underline{\ell}+1,\underline{m}+1)]\underline{p}^{\underline{\ell}+\underline{x}}(1-\underline{p})^{\underline{m}+\underline{n}-\underline{x}}$ (Keats & Lord, 1962, p. 69).

The first integral in Equation 4 can be expressed as follows:

$$\int_0^{\pi} \underline{P}(\underline{x},\underline{p})d\underline{p} = [\binom{\underline{n}}{\underline{x}} / \underline{B}(\underline{\ell}+1,\underline{m}+1)]\int_0^{\pi} \underline{p}^{\underline{\ell}+\underline{x}}(1-\underline{p})^{\underline{m}+\underline{n}-\underline{x}}d\underline{p}$$

$$= [\underline{B}(\underline{\ell}+\underline{x}+1,\underline{m}+\underline{n}-\underline{x}+1)/\underline{B}(\underline{\ell}+\underline{x}+1,\underline{m}+\underline{n}-\underline{x}+1)][\binom{\underline{n}}{\underline{x}} / \underline{B}(\underline{\ell}+1,\underline{m}+1)]\int_0^{\pi} \underline{p}^{\underline{\ell}+\underline{x}}(1-\underline{p})^{\underline{m}+\underline{n}-\underline{x}}d\underline{p}$$

$$= [\underline{B}(\underline{\ell}+\underline{x}+1,\underline{m}+\underline{n}-\underline{x}+1)\binom{\underline{n}}{\underline{x}} / \underline{B}(\underline{\ell}+1,\underline{m}+1)]\int_0^{\pi} [1/\underline{B}(\underline{\ell}+\underline{x}+1,\underline{m}+\underline{n}-\underline{x}+1)]\underline{p}^{\underline{\ell}+\underline{x}}(1-\underline{p})^{\underline{m}+\underline{n}-\underline{x}}d\underline{p}$$

$$= [\underline{B}(\underline{\ell}+\underline{x}+1,\underline{m}+\underline{n}-\underline{x}+1)\binom{\underline{n}}{\underline{x}} / \underline{B}(\underline{\ell}+1,\underline{m}+1)]\underline{I}_{\pi}(\underline{\ell}+\underline{x}+1,\underline{m}+\underline{n}-\underline{x}+1), \tag{5}$$

where $\binom{\underline{n}}{\underline{x}}$, $\underline{B}(\ ,\ )$, and $\underline{I}(\ ,\ )$ are respectively the binomial coefficient, the beta function, and the incomplete beta function.

Similarly the second integral in Equation 4 can be written:

$$\int_{\pi}^{1} \underline{P}(\underline{x},\underline{p})d\underline{p} = [\binom{\underline{n}}{\underline{x}} / \underline{B}(\underline{\ell}+1,\underline{m}+1)]\int_{\pi}^{1} \underline{p}^{\underline{\ell}+\underline{x}}(1-\underline{p})^{\underline{m}+\underline{n}-\underline{x}}d\underline{p}$$

$$= [\underline{B}(\underline{\ell}+\underline{x}+1,\underline{m}+\underline{n}-\underline{x}+1)\binom{\underline{n}}{\underline{x}} / \underline{B}(\underline{\ell}+1,\underline{m}+1)]\int_{\pi}^{1} [1/\underline{B}(\underline{\ell}+\underline{x}+1,\underline{m}+\underline{n}-\underline{x}+1)]\underline{p}^{\underline{\ell}+\underline{x}}(1-\underline{p})^{\underline{m}+\underline{n}-\underline{x}}d\underline{p}$$

$$= [\underline{B}(\underline{\ell}+\underline{x}+1,\underline{m}+\underline{n}-\underline{x}+1)\binom{\underline{n}}{\underline{x}} / \underline{B}(\underline{\ell}+1,\underline{m}+1)][1-\int_0^{\pi} [1/\underline{B}(\underline{\ell}+\underline{x}+1,\underline{m}+\underline{n}-\underline{x}+1)]\underline{p}^{\underline{\ell}+\underline{x}}(1-\underline{p})^{\underline{m}+\underline{n}-\underline{x}}d\underline{p}]$$

$$= [\underline{B}(\underline{\ell}+\underline{x}+1,\underline{m}+\underline{n}-\underline{x}+1)\binom{\underline{n}}{\underline{x}} / \underline{B}(\underline{\ell}+1,\underline{m}+1)][1-\underline{I}_{\pi}(\underline{\ell}+\underline{x}+1,\underline{m}+\underline{n}-\underline{x}+1)]. \tag{6}$$

Finally, Equation 2 is obtained by substituting 5 and 6 into 4:

$$Q = \sum_{x=0}^{c-1} [B(\ell+x+1, m+n-x+1)\binom{n}{x}/B(\ell+1, m+1)]I_{\pi}(\ell+x+1, m+n-x+1) +$$

$$\sum_{x=c}^{n} [B(\ell+x+1, m+n-x+1)\binom{n}{x}/B(\ell+1, m+1)][1-I_{\pi}(\ell+x+1, m+n-x+1)]$$

$$= \{1/B(\ell+1, m+1)\}\{\sum_{x=0}^{c-1}\binom{n}{x}B(\ell+x+1, m+n-x+1)I_{\pi}(\ell+x+1, m+n-x+1) +$$

$$\sum_{x=c}^{n}\binom{n}{x}B(\ell+x+1, m+n-x+1)[1-I_{\pi}(\ell+x+1, m+n-x+1)]\}.$$

References

Griffiths, D. A. Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. Biometrics, 1973, 29, 637-648.

Huynh, H. On consistency of decisions in criterion-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.

Keats, J. A., & Lord, F. M. A theoretical distribution for mental test scores. Psychometrika, 1962, 27, 59-72.

LaValle, I. H. An Introduction to Probability, Decision, and Inference. New York: Holt, Rinehart and Winston, 1970.

Lord, F. M. A strong true score theory, with applications. Psychometrika, 1965, 30, 239-270.

Lord, F. M. Estimating true-score distributions in psychological testing (An empirical Bayes estimation problem). Psychometrika, 1969, 34, 259-299.

Pearson, K. (Ed.) Tables of the Incomplete Beta-Function. New Rochelle, N.Y.: Cambridge University Press, 1956.

Shenton, L. R. Maximum likelihood and the efficiency of the method of moments. Biometrika, 1950, 37, 111-116.

Subkoviak, M. J., & Albrecht, B. A. Empirical investigation of procedures for estimating reliability for mastery tests. Manuscript submitted for publication, 1977.

Wilcox, R. Estimating the likelihood of false-positive and false-negative decisions in mastery testing: An empirical Bayes approach. Journal of Educational Statistics, 1977, in press.

Footnotes

Michael Subkoviak and Rand Wilcox were supported respectively by National Institute of Education Grants No. NIE-G-76-0088 and NIE-G-76-0083. Equal authorship is implied.

[1]Maximum likelihood estimates of $\underline{\ell}$ and $\underline{m}$ are also possible (see Griffiths, 1973). However, when the number of examinees is large, results reported by Shenton (1950) would seem to imply little difference between maximum likelihood estimates and the moment estimates of $\underline{\ell}$ and $\underline{m}$ given here.

Table 1

Values of $\binom{4}{x}$, $B(\ ,\ )$ and $I_{\pi}(\ ,\ )$ for the Numerical Example[a]

x	$\binom{4}{x}$	$B(9.76+x,13.76-x)$	$I_{.85}(9.76+x,13.76-x)$
0	1	$.12\times10^{-6}$	.99
1	4	$.95\times10^{-7}$	.99
2	6	$.87\times10^{-7}$	.99
3	4	$.95\times10^{-7}$	.99
4	1	$.12\times10^{-6}$	.99

[a] $B(9.76,9.76) = .15\times10^{-5}$

ESTIMATING TRUE SCORE IN THE COMPOUND BINOMIAL ERROR MODEL

Rand R. Wilcox
Center for the Study of Evaluation
University of California at Los Angeles

ABSTRACT

Using Monte Carlo techniques, several empirical Bayes procedures for estimating true score in the compound binomial error model are compared. Situations are found where the sample mean performs substantially better in terms of an ensemble squared error loss function. However, such situations would seem to be unlikely in practice. Moreover, the empirical Bayes procedures usually perform as well as the sample mean and frequently they provide a nontrivial improvement.

1. INTRODUCTION

In certain testing situations for which an item sampling model applies, we may want to estimate the percent correct true score of several examinees based on a relatively small number of items. In Individually Prescribed Instruction (IPI), for example, a so-called criterion-referenced or mastery test is administered which typically consists of 25 items or less. It is desired to obtain as accurate an estimate as possible of the examinee's true score; yet the administration of more than a small number of items is frequently impractical.

If the distribution of observed test scores is binomial, the sample mean as an estimate of a single examinee's percent correct true score is maximum likelihood, sufficient and efficient. However, there is reason to believe, as indicated below, that in terms of squared error loss, a substantial improvement over the sample mean can be obtained when several examinees are measured simultaneously. The problem of estimating the true score of several examinees is a special case of the compound statistical decision problem (see Samuel, 1967, for a general exposition). Here we consider the use of empirical Bayes techniques to solve this problem. We note that many of the results for empirical Bayes estimation rules for estimating population means are not presented as compound decision rules. Instead it is assumed that we are given the observed score of k populations (examinees) and the problem is to estimate only the mean of the kth population based on the k observed scores. Of course the

investigator may be interested in accurate estimates of the mean of the other k-1 populations. However, the accuracy of empirical Bayes estimation procedures are frequently described in terms of the kth population only. Thus, we have a slight though important distinction between the present problem and many of the results given for empirical Bayes procedures.

We restrict our attention to dichotomously scored test items. In terms of a single examinee, this suggests that the binomial error model is appropriate (Lord and Novick, 1968, chapter 23). In this special case the sample mean, as noted above, is a strong candidate for estimating an examinee's true score. For the more general problem of estimating the true score of $k > 1$ examinees, the compound binomial error model is more appropriate when each examinee takes the same n randomly sampled items (Lord, 1965; Lord and Novick, 1968). Accordingly, two empirical Bayes estimators of the percent correct true score are described for the more general compound binomial distribution. Of particular interest is the accuracy of these estimators when the number of examinees and the number of items on a test are small or of moderate size. Using Monte Carlo techniques, we also examine the efficacy of using the binomial distribution as a single term approximation to the compound binomial when estimating true score.

2. THE JAMES-STEIN ESTIMATOR

To further motivate the present study, as well as the first estimation procedure, we review briefly the James-Stein estimator for the simplest case. Let y_i, $i=1,\ldots,k$, be k independent and normally distributed random variables with mean $Ey_i = \mu_i$ and variance one. We describe the James-Stein estimator in terms of a single observation from each of the k populations because this formulation is particularly convenient for present purposes. For a slightly more general formulation see Efron (1975) and Efron and Morris (1972b). We use an ensemble squared error loss function, viz.,

$$L(\underline{\mu}, \hat{\underline{\mu}}) = \sum_{i=1}^{k} (\mu_i - \hat{\mu}_i)^2/k \tag{2.1}$$

to measure the accuracy of any statistic $\hat{\underline{\mu}}$ which estimates $\underline{\mu} = (\mu_1,\ldots,\mu_k)$. The risk (expected loss) of the sample mean $\underline{y} = (y_1,\ldots,y_k)$ is

$$R(\underline{\mu}, \hat{\underline{\mu}}) = E\sum_{i=1}^{k} (y_i - \mu_i)^2/k = 1 \tag{2.2}$$

which is a constant for fixed k. The estimation procedure proposed by James and Stein (1961) is to replace the sample mean y_i with

$$y_i^{(1)} = \mu_i' + (1 - (k-2)/S)(y_i - \mu_i'), \quad i=1,\ldots,k \tag{2.3}$$

where $\underline{\mu}' = (\mu_1',\ldots,\mu_k')$ is any initial guess at $\underline{\mu}$ and $S = \sum_{j=1}^{k} (y_j - \mu_j')^2$. The risk of this estimator is less than 1, the risk of the sample mean, when $k \geq 3$. Efron and Morris (1971, 1972a, 1972b, 1973a, 1973b) have examined the James-Stein estimator at some length. All indications are that in terms

of the ensemble loss function (2.1), (2.3) can provide a nontrivial improvement over the sample mean.

As illustrated by Efron and Morris (1975, 1972a) the James-Stein estimator arises in an empirical Bayes context as follows: If μ_i, $i=1,\ldots,k$, are themselves a random sample from a normal prior distribution with mean λ and unknown variance τ^2, the Bayes estimate of $\underline{\mu}$ is the a posteriori mean:

$$E(\mu_i|y_i) = \lambda + (1-(1+\tau^2)^{-1})\ (y_i-\lambda) \tag{2.4}$$

The unknown quantity $(1+\tau^2)^{-1}$ can be estimated since

$$E[(k-2)/S] = (1+\tau^2)^{-1} \tag{2.5}$$

(see Efron and Morris, 1975, expression 1.10). Substituting $(k-2)/S$ for $(1+\tau^2)^{-1}$ in (2.4) results in the James-Stein estimator (2.3).

Let θ_i be the percent correct true score of the ith examinee and let x_i be the corresponding observed score (number correct) on an n item test. Let $h(x|\theta)$ be the conditional probability distribution of observed test scores. When $h(x|\theta)$ is binomial and assuming the test scores are independent among the examinees, we may use the James-Stein estimator as illustrated by Efron and Morris (1975, section 2). In particular, they used an arc-sine transformation for stabilizing the variance of a binomial distribution, viz.,

$$y_i = n^{\frac{1}{2}} \text{ arc } \sin(2x_i/n-1) \tag{2.6}$$

The random variable y_i given by (2.6) is approximately normally distributed with mean $\zeta_i = n^{\frac{1}{2}}$ arc sin $(2\theta_i-1)$ and variance one, the variance being independent of θ_i. Consequently, we may use (2.3) to estimate ζ_i. Efron and

Morris use the James-Stein estimator to estimate ζ_i but they follow Lindley's suggestion (Lindley, 1962) and estimate the unknown value λ, the mean of the assumed normal prior, with $\bar{y} = \Sigma y_i/k$. The estimate of ζ_i is

$$\hat{\zeta}_i = \bar{y} + (1-(k-3)/V)\ (y_i-\bar{y}) \tag{2.7}$$

where $V = \Sigma(y_i-\bar{y})^2$ and where k-3 is the appropriate constant since one parameter is estimated. The resulting estimate of θ_i is

$$\hat{\theta}_i = \tfrac{1}{2}\ [\sin(\hat{\zeta}_i\ n^{-\frac{1}{2}}) + 1] \tag{2.8}$$

Following Lord (1965) and as suggested by Lord and Novick (1968 chapter 23) we use a two-term approximation to the compound binomial given by

$$h(x|\theta) = p_n(x) + d\theta\ (1-\theta)\ C(x) \tag{2.9}$$

where

$$p_n(x) = \binom{n}{x}\ \theta^x\ (1-\theta)^{n-x}$$

$$C(x) = \sum_{v=0}^{2} (-1)^{v+1}\ \binom{2}{v}\ p_{n-2}\ (x-v)$$

The parameter d is given by

$$d = \frac{n^2\ (n-1)\ \sigma_\pi^2}{2[\mu_x\ (n-\mu_x) - \sigma_x^2 - n\sigma_\pi^2]} \tag{2.10}$$

where σ_π^2 is the variance of the item difficulties, and where μ_x and σ_x^2 are the population mean and variance of the observed score distribution. The mean and variance of this approximation to the compound binomial are

(2.11) $E(x|\theta) = n\theta$

(2.12) $\text{Var}(x|\theta) = (n-2d)\,\theta\,(1-\theta)$

(Lord, 1965, p. 265). It follows that

(2.13) $$y = \frac{2n}{\sqrt{n-2d}} \sin^{-1}(\sqrt{x/n})$$

is asymptotically normal with mean $\zeta = 2n \sin^{-1}(\sqrt{\lambda})$ and variance one (Rao, 1973, section 6g). Consequently we may estimate d using standard item analysis techniques and then apply the James-Stein estimator via (2.7) and use $\hat{\theta}_i = \sin^2(\hat{\zeta}_i/2n)$

Before leaving this area we note that the James-Stein estimator has good properties vis-a-vis the ensemble loss function (2.1) but that in terms of individual components of the vector $\underline{\theta}$ it may have risk as large as k/4 as great as the risk of the sample mean. This problem has been investigated by Efron and Morris (1971, 1972a) where limited translation estimators are introduced. We consider this point further in section 5.

3. THE OPTIMAL LINEAR ESTIMATOR

Previous experience with the James-Stein estimator suggests that it may provide a substantial improvement over the sample mean in terms of squared error loss. However, there are several reasons for considering alternate estimation procedures. As noted above, the use of the James-Stein estimator is dependent upon an inverse sine transformation which

is asymptotically normal. As indicated earlier, our main concern is with estimating true score with a relatively small number of items. Consequently, the advantage of the James-Stein estimator may be less apparent. In addition, the rate of convergence to normality for inverse sine transformations is a function of the unknown true score of the examinee. Finally there is the estimation of d which may minimize any advantages of (2.8).

We observe that the binomial analogue of the James-Stein estimator can be derived using empirical Bayes techniques. This can be accomplished by assuming a beta prior and estimating the parameters of the prior by estimating the first and second moments of the true score distribution (Maritz, 1970, p. 22). Even for the more general two-term approximation to the compound binomial it is possible to estimate the parameters of an assumed beta prior using the results given by Lord (1965). However, we do not consider this estimation procedure directly. Instead we apply a result due to Griffin and Krutchkoff (1971) which is more general and mathematically more convenient.

Let $\hat{\theta}_i = \alpha x_i + \beta$ be a decision rule for estimating the percent correct true score of the ith examinee (i=1,...,k). Consider the loss function $(\theta_i - \hat{\theta}_i)^2$. Assuming the first two moments of $h(x|\theta)$ and the true score distribution exist and since $E(x|\theta) = n\theta$, the optimal (minimum risk) linear estimator is given by setting

$$\alpha = \frac{n^2 \, VAR(\theta)}{VAR(x)} \tag{3.1}$$

$$\beta = (1 - \alpha) \, E(\theta) \tag{3.2}$$

This result holds without making any assumption about the form of the prior. Thus, we estimate the true score of the ith examinee to be

$$\hat{\theta}_i = \frac{n^2 \text{ VAR } (\theta)}{\text{VAR}(x)} [x_i - E(\theta)] + E(\theta) \tag{3.3}$$

From (2.11) and (2.12) we have that

$$E(x^2|\theta) = (n-2d)\ (\theta-\theta^2) + n^2\theta^2 \tag{3.4}$$

Therefore,

$$E(x^2) = (n-2d)\ [E(\theta) - E(\theta^2)] + n^2\ E(\theta^2) \tag{3.5}$$

Since $E(x) = EE(x|\theta) = nE(\theta)$ we have that

$$\text{VAR}(x) = (n-2d)\ [E(\theta) - E(\theta^2)] + n^2\ E(\theta^2) - n^2E^2(\theta) \tag{3.6}$$

Writing $\text{VAR}(\theta)$ as $E(\theta^2) - E^2(\theta)$ and substituting the right hand side of (3.6) for $\text{VAR}(x)$ in (3.1) gives

$$\alpha = \frac{n^2[E(\theta^2) - E^2(\theta)]}{(n-2d)\ [E(\theta) - E(\theta^2)] + n^2[E(\theta^2) - E^2(\theta)]} \tag{3.7}$$

Since

$$E_\theta[E(x/n|\theta)] = E(\theta)$$

$$E_\theta[\frac{E(x^2 - (1-2d/n)\ x|\theta)}{n^2 - n + 2d}] = E(\theta^2),$$

where E_θ means expectation over the distribution of θ,

we may estimate $E(\theta)$ and $E(\theta^2)$ with

$$\hat{\mu}_1 = \frac{1}{k} \sum_{i=1}^{k} x_i/n \tag{3.8}$$

$$(3.9) \qquad \hat{\mu}_2 = \frac{1}{k} \sum_{i=1}^{k} \frac{x_i^2 - (1-2d/n)\, x_i}{n^2 - n + 2d}$$

One difficulty with using (3.8) and (3.9) is that the resulting estimate of VAR(θ) as well as VAR(x) may be negative. In this event we follow Griffin and Krutchkoff and estimate α to be one. After estimating d we have an estimate of α and β which yields an estimate of the optimal linear estimator.

Griffin and Krutchkoff (1971, p. 196) note that the ratio of the risks of the optimal linear estimator to the sample mean is

$$(3.10) \qquad Q = \frac{n^2 VAR(\theta)}{VAR(x)} \leq 1$$

It appears that as d gets large the advantage of the Griffin - Krutchkoff estimator becomes negligible since VAR(x) decreases. Moreover, from (3.6) we see that as d approaches n/2, VAR(x) approaches VAR(θ). Hence, Q given by (3.10) approaches one. In practice however the Griffin - Krutchkoff estimator may provide a nontrivial improvement over the sample mean. Lord (1965) notes that the two-term approximation to the compound binomial given by (2.10) is a close approximation to a frequency distribution but that difficulties might arise when d is too large. For all 16 distributions reported by Lord the estimated value of d was in the closed interval [0.0, 4.0]. Consequently it may be realistic and mathematically convenient to assume $0 \leq d \leq 4$. Such a restriction on d suggests that the Griffin - Krutchkoff estimator may be useful even for moderately short tests. We consider this point further in section 5.

4. OTHER ESTIMATION PROCEDURES

Given the observed number of successes in k binomial populations, a number of procedures have been suggested for estimating simultaneously the probability of success in the populations. Since the binomial distribution is a one-term approximation to the compound binomial, these estimation procedures deserve consideration. It may be that altering the value of d in (2.9) has little effect on the resulting estimates. Accordingly, we list the estimators that are included in the Monte Carlo studies below.

1. The first estimator is the James-Stein estimator but with the arc-sine transformation (2.13) replaced with

$$(4.1) \qquad y = \tfrac{1}{2}(4n+2)^{\frac{1}{2}}\left[\sin^{-1}\sqrt{\frac{x}{n+1}} + \sin^{-1}\sqrt{\frac{x+1}{n+1}}\right].$$

(Freeman and Tukey, 1950). Other arc-sine transformations exist but we chose to use (4.1) because Mosteller and Youtz (1961) as well as Mosteller and Tukey (1968) indicate that it is the best existing angular transformation for the binomial case. After applying (4.1) we calculate $\hat{\xi}_i$ using (2.7) and then compute

$$(4.2) \qquad \hat{\theta}_i = \sin^2\left[(4n+2)^{-\frac{1}{2}}\hat{\xi}_i\right]$$

2. The Griffin - Krutchkoff estimator but with d set arbitrarily equal to zero in the estimate of α as given by expression (3.7). When we use the Griffin - Krutchkoff estimator as described in section 3 above, we use the exact value of d, i.e., the value of d which was used in the Monte Carlo studies to generate observed scores via (2.9). By doing so, we will have some indication of the value of having accurate estimates of d. If using the exact value of d in (3.7) has negligible effect in terms of lowering the squared error loss, there would seem to be little to gain in estimating d as opposed to setting d arbitrarily equal to zero.

3. The sample mean $\hat{\theta}_i = X_i/n$

4. Leonard's estimator (Leonard, 1976) adapted to the binomial distribution. We compute

$$\hat{\tau}_i = \rho y_i + (1-\rho)\bar{y} \tag{4.3}$$

where y_i is given by (4.1), $1-\rho = (k-3)/S^2 - G[\frac{1}{2}(k-3), S^2/2]$, $S^2 = \Sigma(y_i-\bar{y})^2$, $\bar{y} = \Sigma y_i/k$, $G[r, t] = y^{r-1} e^{-t}/\gamma(r, t)$ and where $\gamma(r,t) = \int_0^t u^{r-1} e^{-u} du$ is the incomplete gamma function. Once $\hat{\tau}_i$ is calculated, we use (4.2) to estimate the examinee's true score. Leonard's estimator may look formidable but it is easily evaluated on a computer.

A number of estimators were not included in this study because there are known situations where the sample mean provides a substantially better estimate in terms of squared error loss. These were the minimax estimator

(see, e.g., Copas, 1972), Lemon's estimator and Lemon's iterated estimator (Lemon and Krutchkoff, 1969), Copas' first and second estimator (Copas, 1972) and the pooled estimator (George, 1971). We also did not consider the method of Maritz (1970, chapter 3) which uses a step-function approximation of the true score distribution. The difficulty with this procedure is that it requires the minimization of a rather complicated function and so it is expensive to evaluate. (Note that the Monte Carlo results given by Maritz, Table 3.10, p. 69, are based on only 20 iterations.) Subkoviak (1976) mentions still another estimate but he points out that it can lead to erroneous estimates under certain circumstances and so we did not include it in this study.

5. MONTE CARLO RESULTS

All of the estimators described above are known to have good asymptotic properties when estimating the probability of success in k independent binomial populations. The problem is determining when they should be used in terms of the number of examinees tested as well as the number of items on the test. There is also the problem of determining the robustness of the estimators when observed scores arise from a compound binomial distribution.

As a preliminary investigation, we estimated the risk

$$(5.1) \qquad R = E \sum_{i=1}^{k} (\theta_i - \hat{\theta}_i)^2/k,$$

the expected value of the ensemble squared error loss function, for various values of $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k)$. Some of the vectors of true scores we used are reported in Table 1. The corresponding estimated risks are reported in Table 2 for n=5, 10, 20, d=0., 2.5, 4.0. All results were based on 500

iterations. The results in Table 2 serve two useful purposes: 1) they act as a check on the accuracy of the Monte Carlo studies, and 2) they illustrate the potential advantages of the empirical Bayes estimation procedures. Note that we did not consider the case n=5 for d=2.5 or d=4.0. Since the variance of the two-term approximation to the compound binomial is $(n-2d)\theta(1-\theta)$, the variance is less than or equal to zero when n=5, $d \geq 2.5$ and so the compound binomial error model makes little sense.

Observe that when the true scores are clustered together, for example, when all the true scores are .7 (true score vector number 1 and number 8 in Table 2), the empirical Bayes procedures are considerably better than the sample mean. However, no one estimator dominated the other, i.e., always had the smallest risk. With one exception (true score vector number 12) we did not find a situation where the ratio of the risk of the sample mean to the risk of an empirical Bayes procedure was particularly small. The implication is that our estimates of the examinees' true score may be less accurate using say the Griffin - Krutchkoff estimator but only by a slight amount. However, the Griffin - Krutchkoff estimator may be considerably more accurate than the sample mean. This finding is consistent with the Monte Carlo results obtained by Martz and Lian (1974) in an investigation similar to ours.

There remains the question about the effect of the parameter d. From the results in Table 2 we see that the risk of the sample mean decreases, as it should, since as indicated earlier, the variance of the sample mean decreases as d gets large. The interesting finding is that the risk of the

other estimation procedures also tended to decrease suggesting that they may provide good results even when the binomial error model does not hold.

Mixed results were obtained for the estimators described in sections 2 and 3 (those using the exact value of d) when compared to those described in section 4 (d set arbitrarily equal to zero). For the Griffin - Krutchkoff estimator we found that using the exact d in (3.7) tended to give slightly worse results when observed scores were generated with d=2.5. When scores were generated with d=4.0, using the exact value of d in (3.7) tended to improve our estimates of the true scores. Apparently, using the Griffin - Krutchkoff estimator as described in section 3 begins to make a difference only for extreme values of d (i.e. values of d close to 4.0). Of course the actual value of d is not known in practice and so there is the additional problem of determining the effects of estimating d. As for the James-Stein estimator, again we found that for d=4.0, the exact value of d improved our estimates in some but not all instances.

Based on our preliminary investigation, Leonard's estimator seems to be the most attractive. As indicated in Table 2, there are instances where the sample mean performs better. However, Leonard's estimator usually performed at least as well as the sample mean and frequently it performed much better, even for $d \neq 0$. Consider, for example true score vector number 8, d=4.0, n=10. The risk of the sample mean was estimated to be .0082 while for Leonard's estimator it is .0018. This is a substantial improvement. In this particular case the James-Stein estimator also did well. There are, however, instances where Leonard's estimator did well but the James-Stein

estimator did not. For example, there is true score vector number 1, d=4.0, n=10.

Another interesting finding is that the advantages of the empirical Bayes procedures appear to be a function of the ratio k/n. As n gets large for fixed k, our findings suggest that the difference between all of the estimators becomes small.

As indicated earlier, the James-Stein estimator has good properties vis-a-vis (5.1) but it may be considerably less accurate for a particular true score being estimated. That is, for some specific θ_i, the James-Stein estimator may be less accurate, on the average, than the sample mean despite the fact that the James-Stein estimator performs better when judged via an ensemble squared error loss function. We decided, therefore, to also estimate the expected value of $t=\max_i \{(\theta_i-\hat{\theta})^2\}$. In otherwords, we are interested in the average value of the maximum squared error component loss, the maximum being taken over all k true scores being estimated. The results are reported in Table 3. As can be seen, the sample mean performed satisfactorily for d=4.0 but it did rather poorly for d=0.0, 2.5.

One could argue that there really are configurations of true scores where the sample mean gives a non-trivial improvement over the other procedures described in section 4 but that we failed to find it. One approach to this problem is to consider the Bayes risk which is the expected value of (5.1) over $g(\theta)$, the distribution of true scores. We may write the Bayes risk as

$$R_B = E_\theta \, E \frac{1}{k} \Sigma(\theta_i-\hat{\theta}_i)^2 \tag{5.2}$$

Again using Monte Carlo procedures, we evaluated (5.2) for a number of true score distributions. This was accomplished by generating true scores according to a beta distribution and then generating observed scores according to (2.9). More precisely, k true scores were generated assuming

$$(5.3) \qquad g(\theta) = \frac{\Gamma(r+s)}{\Gamma(r)\,\Gamma(s)}\,\theta^{r-1}\,(1-\theta)^{s-1},$$

where Γ is the usual gamma function. Expression (5.3) is motivated as the natural conjugate prior of the binomial distribution (Raiffa and Schlaifer, 1961) as well as from empirical findings that (5.3) provides a good fit to true score distributions (Keats and Lord, 1962). The parameters r and s were chosen so as to include a variety of true score distributions, namely, J-shaped, symmetric, skewed, unimodal, and multimodal priors. The multimodal true score distributions were formed with mixtures of beta distributions, i.e., we formed the random variable $\theta = c_1\,\xi_1 + c_2\xi_2$ where ξ_1 and ξ_2 are two beta distributions and c_1 and c_2 are positive constants such that $c_1+c_2=1$.

Consistent with earlier investigations, we found the improvement of the empirical Bayes estimation procedures over the sample mean to be sensitive to the variance of the true score distribution. More specifically, a decrease in the variance of the true score distribution tended to increase the ratio of the risk of the sample mean to the risk of the empirical Bayes estimates.

Table 4 gives the estimated Bayes risk (the estimate of R_B) for (r, s) = (1, 1), (2, 2), (3, 3), (4, 4), (5, 5), d=0., 2.5, 4.0. We chose to

report our results for these values of r and s because they illustrate the effect of the variance of the true score distribution on the value of R_B. When r=s, the true score distribution is symmetric about the point θ=.5. As r (=s) increases, the variance decreases. Table 4 illustrates that as the true score variance decreases, the improvement of the empirical Bayes estimators over the sample mean (in terms of the ratio of the risks) increases. Consider for example, n=5, r=s=1, d=0. The ratio of the risk of the sample mean to Leonard's estimator is .0427/.0337=1.27. For r=s=5 the ratio is .0449/.0271=1.66. Of course our empirical findings do not prove that the ratio of the risk of the sample mean to the risk of an empirical Bayes estimator will increase as the true score variance decreases. However, our results together with earlier investigations suggest that this will generally be the case. The finding that is of particular interest here is that this result seems to hold even when the observed scores arise from a two-term approximation to the compound binomial distribution.

THE BASEBALL PROBLEM REVISITED

To illustrate the potential advantages of the James-Stein estimator, Efron and Morris (1975) considered the problem of predicting the 1970 baseball season batting averages of 18 major league baseball players based on their first 45 official at bats for the 1970 season. As a final demonstration of the estimation procedures considered here, we reestimated the batting averages of the same 18 players using the data given in Table 1 of

Efron and Morris (1975). The results are given in Table 5. The first column headed θ_i gives the batting average of the 18 players at the end of the season. We set d=0 in this particular instance and so it was necessary to report only one estimate using the Griffin-Krutchkoff estimation technique.

The corresponding loss of the five estimators used in Table 5, as defined by (2.1), were calculated to be .00419, .00126, .001196, .001199, .00126, respectively. The ratio of the loss of the sample mean to the loss of another estimator serves as a measure of the efficiency of the estimator. For the Griffin-Krutchkoff estimator, for example, this ratio is .00419/.00126=3.33. Based on this measure of efficiency, all four empirical Bayes estimators are substantially better than the sample mean in terms of squared error loss.

We observe that the choice of a particular arc-sine transformation had little effect on the James-Stein estimator. The Freeman-Tukey transformation (4.1) is known to be particularly useful when n, the test length, is small. Even for moderate size n, the advantage of (4.1) appears to be negligible when estimating percent correct true score via (2.7).

It is interesting to compare the sample mean to Leonard's estimator for each of the θ_i's. We find that the sample mean is more accurate for only 3 of the 18 players. Moreover, for one of these three instances, namely θ_i=.210, the difference between the two estimators is trivial. For the Griffin-Krutchkoff estimator the sample mean is more accurate in only four instances and for the James-Stein estimator using the Freeman-Tukey transformation the number is 5.

SUMMARY AND CONCLUDING REMARKS

Several empirical Bayes estimation procedures have been described for estimating percent correct true score in the compound binomial error model. Our extensive empirical investigations suggest that the sample mean is poorest in terms of an ensemble squared error loss function. This does not mean that there are no situations where the sample mean will give more accurate results. However, it appears that there is usually little to lose and frequently a great deal to gain by using empirical Bayes procedures. Moreover, our findings suggest that these estimation techniques are robust in terms of departure from the binomial distribution. We also found that estimation procedures derived under the more general two-term approximation to the compound binomial error model appear to offer little advantage over estimators intended for the simpler binomial distribution.

Confronted with a situation where the goal is to estimate the true score of $k \geq 5$ examinees and where a squared error loss function is appropriate, our findings suggest that Leonard's estimator is best with the Griffin-Krutchkoff estimator a close second. Our reason for excluding the James-Stein estimator is that on rare occasions it did rather poorly compared to the sample mean but Leonard's estimator and the Griffin-Krutchkoff estimator did rather well (see, e.g., the results in Table 2 for true score vector number 6, d=0., n=20). There are situations where the Griffin-Krutchkoff estimator and the sample mean perform substantially better than Leonard's estimator as evidenced by the results in Table 2 for true score vector number 12. However, such a configuration of true scores would seem to be highly unlikely in most testing situations.

Table 1

Vectors of True Scores Used in Monte Carlo Studies

1. .7, .7, .7, .7, .7
2. .6, .7, .8, .85, .9
3. .1, .2, .65, .79, .81
4. .1, .8, .83, .85, .9
5. .1, .15, .2, .9, .9
6. .9, .91, .92, .93, .94
7. .1, .3, .5, .8, .95
8. .7, .7, .7, .7, .7, .7, .7, .7, .7, .7
9. .6, .65, .7, .77, .79, .83, .86, .89, .91, .95
10. .2, .8, .9, .91, .92, .95, .98, .99, .99, 1.0
11. .1, .2, .3, .4, .5, .6, .7, .8, .9, .95
12. .05, .05, .05, .05, .05, .95, .95, .95, .95, .95
13. .8, .8, .9, .91, .92, .95, .98, .99, .99, 1.0

Table 2 (continued)

AVERAGE SQUARED ERROR LOSS FOR THE VECTORS
OF TRUE SCORES LISTED IN TABLE 1

True Score Vector	d	N	Sample Mean	Griffin Krutchkoff (using d=0)	Griffin Krutchkoff (using exact d)	James Stein (using exact d)	James Stein (using d=0)	Leonard
8	0	5	.0411	.0224	.0224	.0221	.0116	.0112
8	0	10	.0211	.0125	.0125	.0084	.0064	.0061
8	0	20	.0105	.0055	.0055	.0037	.0034	.0030
9	0	5	.0307	.0167	.0167	.0205	.0138	.0130
9	0	10	.0153	.0097	.0097	.0116	.0099	.0091
9	0	20	.0073	.0060	.0060	.0061	.0058	.0056
10	0	5	.0770	.0603	.0603	.0596	.0504	.0505
10	0	10	.0876	.0787	.0787	.0788	.0714	.0714
10	0	20	.0935	.0891	.0891	.0898	.0854	.0854
11	0	5	.0346	.0301	.0301	.0329	.0279	.0276
11	0	10	.0167	.0157	.0157	.0162	.0154	.0154
11	0	20	.0086	.0083	.0083	.0084	.0082	.0082
12	0	5	.0095	.0100	.0100	.0110	.0148	.0148
12	0	10	.0048	.0050	.0050	.0056	.0073	.0073
12	0	20	.0025	.0025	.0025	.0027	.0033	.0033
13	0	5	.0770	.0481	.0481	.0443	.0311	.0324
13	0	10	.0876	.0731	.0731	.0727	.0631	.0631
13	0	20	.0937	.0868	.0868	.0878	.0827	.0827

Table 2 (continued)

AVERAGE SQUARED ERROR LOSS FOR THE VECTORS
OF TRUE SCORES LISTED IN TABLE 1

True Score Vector	d	N	Sample Mean	Griffin Krutchkoff (using d=0)	Griffin Krutchkoff (using exact d)	James Stein (using exact d)	James Stein (using d=0)	Leonard
1	2.5	10	.0114	.0105	.0120	.0089	.0073	.0051
1	2.5	20	.0080	.0060	.0077	.0045	.0042	.0041
2	2.5	10	.0095	.0079	.0120	.0088	.0099	.0073
2	2.5	20	.0063	.0061	.0075	.0059	.0057	.0057
3	2.5	10	.0098	.0131	.0108	.0096	.0113	.0102
3	2.5	20	.0063	.0069	.0074	.0066	.0064	.0060
4	2.5	10	.0072	.0088	.0095	.0077	.0090	.0091
4	2.5	20	.0048	.0051	.0056	.0045	.0051	.0049
5	2.5	10	.0063	.0075	.0079	.0068	.0084	.0083
5	2.5	20	.0039	.0042	.0046	.0038	.0044	.0045
6	2.5	10	.0049	.0043	.0076	.0155	.0035	.0035
6	2.5	20	.0027	.0020	.0044	.0027	.0052	.0020
7	2.5	10	.0088	.0107	.0102	.0088	.0095	.0097
7	2.5	20	.0057	.0060	.0068	.0058	.0058	.0059

Table 2 (continued)

AVERAGE SQUARED ERROR LOSS FOR THE VECTORS
OF TRUE SCORES LISTED IN TABLE 1

True Score Vector	d	N	Sample Mean	Griffin Krutchkoff (using d=0)	Griffin Krutchkoff (using exact d)	James Stein (using exact d)	James Stein (using d=0)	Leonard
8	2.5	10	.0123	.0116	.0119	.0038	.0080	.0024
8	2.5	20	.0079	.0062	.0081	.0025	.0028	.0018
9	2.5	10	.0090	.0075	.0111	.0073	.0090	.0067
9	2.5	20	.0056	.0053	.0072	.0048	.0049	.0048
10	2.5	10	.0853	.0761	.0864	.0809	.0696	.0696
10	2.5	20	.0928	.0883	.0936	.0900	.0848	.0848
11	2.5	10	.0097	.0113	.0114	.0096	.0110	.0110
11	2.5	20	.0063	.0063	.0075	.0063	.0063	.0063
12	2.5	10	.0036	.0039	.0048	.0038	.0067	.0067
12	2.5	20	.0019	.0020	.0027	.0020	.0028	.0028
13	2.5	10	.0853	.0696	.0865	.0780	.0611	.0613
13	2.5	20	.0929	.0857	.0937	.0884	.0820	.0819

Table 2 (continued)

AVERAGE SQUARED ERROR LOSS FOR THE VECTORS
OF TRUE SCORES LISTED IN TABLE 1

True Score Vector	d	N	Sample Mean	Griffin Krutchkoff (using d=0)	Griffin Krutchkoff (using exact d)	James Stein (using exact d)	James Stein (using d=0)	Leonard
1	4.0	10	.0086	.0086	.0045	.0063	.0087	.0042
1	4.0	20	.0064	.0056	.0041	.0036	.0039	.0032
2	4.0	10	.0072	.0067	.0064	.0070	.0091	.0062
2	4.0	20	.0052	.0056	.0052	.0051	.0053	.0047
3	4.0	10	.0070	.0110	.0072	.0071	.0092	.0092
3	4.0	20	.0051	.0057	.0053	.0052	.0053	.0053
4	4.0	10	.0057	.0085	.0058	.0058	.0084	.0084
4	4.0	20	.0038	.0042	.0039	.0039	.0043	.0043
5	4.0	10	.0051	.0069	.0052	.0052	.0078	.0078
5	4.0	20	.0036	.0039	.0037	.0037	.0041	.0041
6	4.0	10	.0044	.0041	.0026	.0038	.0038	.0038
6	4.0	20	.0026	.0020	.0015	.0019	.0054	.0018
7	4.0	10	.0069	.0097	.0070	.0070	.0084	.0084
7	4.0	20	.0049	.0053	.0050	.0049	.0051	.0051

Table 2 (continued)

AVERAGE SQUARED ERROR LOSS FOR THE VECTORS
OF TRUE SCORES LISTED IN TABLE 1

True Score Vector	d	N	Sample Mean	Griffin Krutchkoff (using d=0)	Griffin Krutchkoff (using exact d)	James Stein (using exact d)	James Stein (using d=0)	Leonard
8	4.0	10	.0082	.0082	.0033	.0038	.0178	.0018
8	4.0	20	.0065	.0060	.0037	.0018	.0036	.0014
9	4.0	10	.0068	.0071	.0056	.0060	.0104	.0066
9	4.0	20	.0048	.0049	.0041	.0041	.0045	.0042
10	4.0	10	.0844	.0747	.0825	.0826	.0689	.0689
10	4.0	20	.0924	.0879	.0897	.0902	.0844	.0844
11	4.0	10	.0074	.0106	.0073	.0074	.0103	.0103
11	4.0	20	.0057	.0060	.0057	.0057	.0060	.0060
12	4.0	10	.0031	.0034	.0031	.0032	.0065	.0065
12	4.0	20	.0017	.0017	.0017	.0018	.0026	.0026
13	4.0	10	.0846	.0685	.0815	.0818	.0608	.0609
13	4.0	20	.0923	.0853	.0882	.0889	.0815	.0815

Table 3

AVERAGE MAXIMUM SQUARED ERROR COMPONENT LOSS
FOR THE VECTORS OF TRUE SCORES LISTED IN TABLE 1

True Score Vector	d'	N	Sample Mean	Griffin Krutchkoff (using d=0)	Griffin Krutchkoff (using exact d)	James Stein (using exact d)	James Stein (using d=0)	Leonard
1	0	5	.0978	.0612	.0612	.0816	.0480	.0506
1	0	10	.0563	.0349	.0349	.0348	.0282	.0307
1	0	20	.0276	.0182	.0182	.0161	.0148	.0151
2	0	5	.0947	.0613	.0613	.0864	.0607	.0588
2	0	10	.0424	.0307	.0307	.0379	.0333	.0306
2	0	20	.0288	.0198	.0198	.0207	.0196	.0188
3	0	5	.0857	.0867	.0867	.0879	.0790	.0779
3	0	10	.0462	.0475	.0475	.0470	.0447	.0447
3	0	20	.0227	.0235	.0235	.0232	.0226	.0226
4	0	5	.0677	.0817	.0817	.0686	.0724	.0707
4	0	10	.0370	.0376	.0376	.0342	.0355	.0354
4	0	20	.0172	.0184	.0184	.0179	.0184	.0184
5	0	5	.0573	.0656	.0656	.0611	.0652	.0651
5	0	10	.0322	.0345	.0345	.0341	.0355	.0355
5	0	20	.0149	.0160	.0160	.0157	.0164	.0164
6	0	5	.0437	.0258	.0258	.1454	.0168	.0252
6	0	10	.0212	.0135	.0135	.0151	.0110	.0132
6	0	20	.0102	.0062	.0062	.0085	.0100	.0066
7	0	5	.0976	.0942	.0942	.0980	.0830	.0824
7	0	10	.0437	.0448	.0448	.0439	.0409	.0409
7	0	20	.0220	.0225	.0225	.0222	.0215	.0215

Table 3 (continued)

AVERAGE MAXIMUM SQUARED ERROR COMPONENT LOSS
FOR THE VECTORS OF TRUE SCORES LISTED IN TABLE 1

True Score Vector	d	N	Sample Mean	Griffin Krutchkoff (using d=0)	Griffin Krutchkoff (using exact d)	James Stein (using exact d)	James Stein (using d=0)	Leonard
8	0	5	.1378	.0732	.0734	.0200	.0341	.0366
8	0	10	.0753	.0430	.0430	.0311	.0220	.0219
8	0	20	.0389	.0201	.0200	.0134	.0119	.0108
9	0	5	.1234	.0644	.0644	.0813	.0535	.0513
9	0	10	.0623	.0372	.0372	.0452	.0371	.0344
9	0	20	.0293	.0226	.0226	.0238	.0217	.0212
10	0	5	.6399	.4836	.4836	.4836	.3614	.3625
10	0	10	.8099	.7247	.7247	.7222	.6414	.6414
10	0	20	.9024	.8591	.8591	.8654	.8195	.8195
11	0	5	.1386	.1122	.1122	.1303	.1063	.1059
11	0	10	.0647	.0593	.0593	.0628	.0578	.0578
11	0	20	.0357	.0339	.0339	.0346	.0333	.0333
12	0	5	.0437	.0486	.0486	.0505	.0687	.0681
12	0	10	.0198	.0220	.0220	.0229	.0317	.0317
12	0	20	.0094	.0103	.0102	.0106	.0140	.0140
13	0	5	.6399	.3959	.3959	.3579	.2349	.2472
13	0	10	.8099	.6755	.6755	.6717	.5759	.5759
13	0	20	.9024	.8359	.8359	.8465	.7950	.7950

Table 3 (continued)

AVERAGE MAXIMUM SQUARED ERROR COMPONENT LOSS
FOR THE VECTORS OF TRUE SCORES LISTED IN TABLE 1

True Score Vector	d	N	Sample Mean	Griffin Krutchkoff (using d=0)	Griffin Krutchkoff (using exact d)	James Stein (using exact d)	James Stein (using d=0)	Leonard
1	2.5	10	.0283	.0264	.0305	.0189	.0192	.0124
1	2.5	20	.0213	.0161	.0207	.0115	.0105	.0106
2	2.5	10	.0250	.0207	.0313	.0244	.0249	.0191
2	2.5	20	.0172	.0170	.0209	.0167	.0159	.0157
3	2.5	10	.0264	.0322	.0294	.0259	.0289	.0269
3	2.5	20	.0174	.0186	.0199	.0179	.0178	.0165
4	2.5	10	.0189	.0247	.0248	.0207	.0240	.0243
4	2.5	20	.0131	.0141	.0151	.0122	.0143	.0137
5	2.5	10	.0167	.0200	.0207	.0186	.0219	.0219
5	2.5	20	.0109	.0118	.0122	.0102	.0123	.0131
6	2.5	10	.0126	.0113	.0183	.0217	.0072	.0084
6	2.5	20	.0073	.0051	.0108	.0062	.0086	.0051
7	2.5	10	.0238	.0281	.0279	.0237	.0254	.0097
7	2.5	20	.0164	.0169	.0193	.0168	.0162	.0166

Table 3 (continued)

AVERAGE MAXIMUM SQUARED ERROR COMPONENT LOSS
FOR THE VECTORS OF TRUE SCORES LISTED IN TABLE 1

True Score Vector	d	N	Sample Mean	Griffin Krutchkoff (using d=0)	Griffin Krutchkoff (using exact d)	James Stein (using exact d)	James Stein (using d=0)	Leonard
8	2.5	10	.0409	.0388	.0396	.0119	.0291	.0069
8	2.5	20	.0281	.0224	.0290	.0080	.0093	.0061
9	2.5	10	.0336	.0260	.0385	.0261	.0287	.0225
9	2.5	20	.0230	.0194	.0268	.0190	.0177	.0176
10	2.5	10	.8099	.7166	.8006	.7647	.6363	.6363
10	2.5	20	.9024	.8580	.8987	.8746	.8188	.8188
11	2.5	10	.0335	.0368	.0406	.0344	.0362	.0361
11	2.5	20	.0246	.0243	.0296	.0251	.0240	.0240
12	2.5	10	.0107	.0127	.0136	.0116	.0212	.0212
12	2.5	20	.0061	.0068	.0084	.0068	.0103	.0103
13	2.5	10	.8099	.6581	.8005	.7400	.5668	.5691
13	2.5	20	.9024	.8325	.8987	.8607	.7939	.7936

Table 3 (continued)

AVERAGE MAXIMUM SQUARED ERROR COMPONENT LOSS
FOR THE VECTORS OF TRUE SCORES LISTED IN TABLE 1

True Score Vector	d	N	Sample Mean	Griffin Krutchkoff (using d=0)	Griffin Krutchkoff (using exact d)	James Stein (using exact d)	James Stein (using d=0)	Leonard
1	4.0	10	.0209	.0208	.0109	.0142	.0218	.0100
1	4.0	20	.0167	.0149	.0106	.0089	.0101	.0080
2	4.0	10	.0181	.0169	.0169	.0175	.0227	.0158
2	4.0	20	.0139	.0154	.0146	.0140	.0143	.0127
3	4.0	10	.0185	.0272	.0192	.0188	.0238	.0237
3	4.0	20	.0138	.0154	.0145	.0141	.0147	.0147
4	4.0	10	.0151	.0239	.0157	.0155	.0223	.0223
4	4.0	20	.0106	.0116	.0110	.0108	.0119	.0119
5	4.0	10	.0126	.0180	.0134	.0131	.0200	.0200
5	4.0	20	.0094	.0106	.0101	.0099	.0112	.0112
6	4.0	10	.0117	.0112	.0070	.0093	.0078	.0084
6	4.0	20	.0070	.0055	.0042	.0049	.0090	.0045
7	4.0	10	.0179	.0241	.0185	.0182	.0214	.0214
7	4.0	20	.0134	.0144	.0139	.0136	.0138	.0138

Table 3 (continued)

AVERAGE MAXIMUM SQUARED ERROR COMPONENT LOSS
FOR THE VECTORS OF TRUE SCORES LISTED IN TABLE 1

True Score Vector	d'	N	Sample Mean	Griffin Krutchkoff (using d=0)	Griffin Krutchkoff (using exact d)	James Stein (using exact d)	James Stein (using d=0)	Leonard
8	4.0	10	.0255	.0255	.0096	.0116	.0637	.0046
8	4.0	20	.0222	.0205	.0128	.0054	.0121	.0041
9	4.0	10	.0251	.0232	.0203	.0214	.0319	.0208
9	4.0	20	.0183	.0167	.0149	.0152	.0149	.0143
10	4.0	10	.8099	.7096	.7910	.7918	.6329	.6329
10	4.0	20	.0902	.8572	.8757	.8802	.8185	.8185
11	4.0	10	.0250	.0330	.0256	.0254	.0327	.0327
11	4.0	20	.0220	.0217	.0215	.0216	.0214	.0214
12	4.0	10	.0076	.0095	.0080	.0081	.0173	.0173
12	4.0	20	.0051	.0057	.0055	.0057	.0091	.0091
13	4.0	10	.8099	.6527	.7797	.7824	.5655	.5655
13	4.0	20	.9024	.8340	.8619	.8694	.7935	.7935

Table 4

AVERAGE SQUARED ERROR LOSS FOR VARIOUS TRUE SCORE DISTRIBUTIONS

(r,s)	d	N	Sample Mean	Griffin Krutchkoff (using d=0)	Griffin Krutchkoff (using exact d)	James Stein (using exact d)	James Stein (using d=0)	Leonard
(1,1)	0	5	.0427	.0372	.0371	.0416	.0346	.0337
(1,1)	0	10	.0288	.0272	.0272	.0283	.0263	.0261
(1,1)	0	20	.0165	.0159	.0159	.0163	.0157	.0156
(2,2)	0	5	.0383	.0299	.0299	.0362	.0289	.0270
(2,2)	0	10	.0201	.0176	.0176	.0188	.0169	.0162
(2,2)	0	20	.0100	.0097	.0097	.0096	.0091	.0089
(3,3)	0	5	.0423	.0303	.0303	.0379	.0299	.0278
(3,3)	0	10	.0205	.0173	.0173	.0189	.0172	.0158
(3,3)	0	20	.0106	.0096	.0096	.0096	.0090	.0089
(4,4)	0	5	.0432	.0296	.0296	.0379	.0296	.0271
(4,4)	0	10	.0218	.0171	.0171	.0198	.0176	.0158
(4,4)	0	20	.0108	.0094	.0094	.0097	.0091	.0088
(5,5)	0	5	.0449	.0294	.0289	.0384	.0293	.0271
(5,5)	0	10	.0226	.0169	.0169	.0182	.0160	.0155
(5,5)	0	20	.0116	.0093	.0093	.0102	.0095	.0091

Table 4 (continued)

AVERAGE SQUARED ERROR LOSS FOR VARIOUS TRUE SCORE DISTRIBUTIONS

(r,s)	d	N	Sample Mean	Griffin Krutchkoff (using d=0)	Griffin Krutchkoff (using exact d)	James Stein (using exact d)	James Stein (using d=0)	Leonard
(1,1)	2.5	10	.0206	.0217	.0200	.0204	.0199	.0197
(1,1)	2.5	20	.0169	.0169	.0167	.0170	.0166	.0165
(2,2)	2.5	10	.0116	.0134	.0110	.0118	.0116	.0104
(2,2)	2.5	20	.0077	.0083	.0077	.0074	.0073	.0071
(3,3)	2.5	10	.0121	.0129	.0109	.0111	.0118	.0099
(3,3)	2.5	20	.0079	.0080	.0075	.0074	.0072	.0069
(4,4)	2.5	10	.0124	.0126	.0105	.0109	.0112	.0093
(4,4)	2.5	20	.0084	.0080	.0076	.0077	.0074	.0069
(5,5)	2.5	10	.0122	.0112	.0100	.0109	.0119	.0088
(5,5)	2.5	20	.0084	.0079	.0074	.0077	.0075	.0069

Table 5

Estimates of the Batting Average of 18 Major League Players Based on Their First 45 at Bats

θ_i	Sample Mean	Griffin Krutchkoff	James Stein using (2.13)	James Stein using (4.1)	Leonard
.346	.400	.273	.290	.290	.308
.298	.378	.272	.286	.287	.302
.276	.356	.270	.282	.283	.296
.222	.333	.269	.277	.280	.289
.273	.311	.268	.273	.276	.282
.270	.311	.268	.273	.276	.282
.263	.289	.267	.268	.272	.276
.210	.267	.265	.264	.269	.269
.269	.244	.264	.259	.265	.262
.230	.244	.264	.259	.265	.262
.264	.222	.263	.254	.261	.255
.256	.222	.263	.254	.261	.255
.303	.222	.263	.254	.261	.255
.264	.222	.263	.254	.261	.255
.226	.222	.263	.254	.261	.255
.285	.200	.262	.249	.257	.248
.316	.178	.260	.244	.253	.241
.200	.156	.259	.239	.249	.233

REFERENCES

Copas, J. B. Empirical Bayes methods and the repeated use of a standard. Biometrika, 1972, 59, 349-360.

Efron, B. Biased versus unbiased estimation. Advances in Mathematics, 1975, 16, 259-277.

Efron, B. and Morris, C. Limiting the risk of Bayes and empirical Bayes estimators - Part I: The Bayes case. Journal of the American Statistical Association, 1971, 66, 807-815.

Efron, B. and Morris, C. Limiting the risk of Bayes and empirical Bayes estimators - Part II: The empirical Bayes case. Journal of the American Statistical Association, 1972a, 67, 130-139.

Efron, B. and Morris, C. Empirical Bayes on vector observations - An extension of Stein's method. Biometrika, 1972b, 59, 335-347.

Efron, B. and Morris, C. Stein's estimation rule and its competitors. Journal of the American Statistical Association, 1973a, 68, 117-130.

Efron, B. and Morris, C. Combining possibly related estimation problems. Journal of the Royal Statistical Society, 1973b, Series B, 35, 379-421.

Efron, B. and Morris, C. Data analysis using Stein's estimator and its generalizations. Journal of the American Statistical Association, 1975, 70, 311-319.

Freeman, M. F. and Tukey, J. W. Transformations related to the angular and the square root. The Annals of Mathematical Statistics, 1950, 21, 607-611.

George, S. L. Evaluation of empirical Bayes estimators for small numbers of past samples. Biometrika, 1971, 58, 244.

Griffin, B. S. and Krutchkoff, R. G. Optimal linear estimators: an empirical Bayes version with application to the binomial distribution. Biometrika, 1971, 58, 195-201.

James, W. and Stein C. Estimation with quadratic loss. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, Berkeley: University of California Press, 1961, 361-379.

Lemon, G. H. and Krutchkoff, R. G. An empirical Bayes smoothing technique Biometrika, 1969, 56, 361-365.

Leonard, T. Some alternative approaches to multiparameter estimation. Biometrika, 1976, 63, 69-75.

Lindley, D. Discussion of paper by C. Stein. Journal of the Royal Statistical Society 1962, Series B, 24, 265-296.

Lord, F. M. A strong true-score theory, with applications. Psychometrika, 1965, 30, 239-270.

Lord, F. M. and Novick, M. R. Statistical Theories of Mental Test Scores. Reading, Mass: Addison - Wesley, 1968.

Maritz, J. S. Empirical Bayes Methods. London: Methuen, 1970.

Martz, H. F. and Lian, M. G. Empirical Bayes estimation of the binomial parameter. Biometrika, 1974, 61, 517-523.

Mosteller, F. and Tukey, J. W. Data analysis, including statistics. In G. Lindzey and E. Aronsen (Eds.). The Handbook of Social Psychology, Reading, Mass: Addison - Wesley, 1968.

Mosteller, F. and Youtz, C. Tables of the Freeman - Tukey transformations for the binomial and Poisson distributions. Biometrika, 1961, 48, 433-440.

Rao, C. R. Linear statistical inference and its applications. New York: John Wiley, 1973.

Samuel, Ester. The compound statistical decision problem. Sankhya, 1967, Series A, 29, 123-140.

Subkoviak, M. Estimating reliability from a single administration of a mastery test. Journal of Educational Measurement, 1976, 13, 265-276.

COMPARING EXAMINEES TO A CONTROL

Rand R. Wilcox
Center for the Study of Evaluation
University of California at Los Angeles

May, 1977

ABSTRACT

When comparing examinees to a control, the examiner usually does not know the probability of correctly classifying the examinees based on the number of items used and the number of examinees tested. Using ranking and selection techniques, a general framework is described for deriving a lower bound on this probability. We illustrate how our general solution can be applied to the binomial error model. New exact results are given for normal populations having unknown and unequal variances.

COMPARING EXAMINEES TO A CONTROL

Setting standards for individuals, whether it be in terms of mental ability, achievement level, or some other characteristic, continues to be a most difficult task. Yet the need for setting standards is ever present. State legislatures have called for the setting of minimum standards for graduation from high school. There is the problem of certifying or licensing individuals for a specific profession such as law, medicine, etc.

One approach to this problem has been suggested by Huynh (1976) in the context of a mastery test. In particular, he suggests using a test to predict success or failure on a suitably chosen external criterion. That is, we administer a test to an examinee and if the observed score is above a predetermined level, we predict that the examinee will succeed at some later task; otherwise we predict failure. (For alternative decision rules and procedures in binary classification, see e.g., Bahadur, 1961; John, 1961; Ott and Kronmal, 1976; Aitchison and Aitken, 1976.)

The purpose of this paper is to suggest an alternative to the approach taken by Huynh. Rather than measure each examinee on some external variable, we compare the examinee to some suitably chosen control, the control being perhaps another examinee (or population of examinees) who has some desired characteristic. For example, we might want to know whether an examinee's reading ability is within a specified range of the average tenth grade student before we award him a high school diploma. A college may want to admit those and only those students whose mathematics achievement is within a specified

range of the average individual graduating in the top one-third of his class. We certainly make no claim that our approach is "better" than the approach taken by Huynh. However, comparing examinees to a control may be more convenient and perhaps more meaningful in certain situations.

The problem we are about to consider belongs to a class of multi-decision statistical techniques known as ranking and selection. To derive a general solution we make the following assumptions:

1. Based on a suitably constructed test we have an observed score x_i $(i=1,\ldots,k)$ for each of k examinees as well as an observed score, x_o, for the control.

2. The probability of obtaining a particular observed score on the ith examinee is a function of the unknown parameter θ_i. θ_o is the corresponding parameter for the control.

3. $\Psi(x_i)$ is a statistic for estimating θ_i.

4. $\Psi(x_i)$ is independent of $\Psi(x_j)$, $i \neq j$.

5. $F(x;\theta_i)$ is stochastically increasing in θ_i, $i=1,\ldots,k$.

That is, $\theta < \theta'$ implies $F(x;\theta') \leq F(x;\theta)$. If, for example, θ is a location parameter, $F(x;\theta)$ is stochastically increasing in θ. No assumption is made concerning item homogeneity, item difficulty, test reliability, etc.

Let $\delta^* > 0$ be a preassigned constant. Based on the statistics $\Psi(x_o)$, $\Psi(x_1),\ldots,\Psi(x_k)$ each examinee is assigned to one of two disjoint sets, say S_B and S_G. A correct decision is made if every examinee having $\theta_i \geq \theta_o + \delta^*$ is assigned to the group S_G and simultaneously every examinee having

$\theta_i \leq \theta_o - \delta^*$ is assigned to S_B. The goal is to make a correct decision (CD) with probability at least P^*, $2^{-k} < P^* < 1$. That is, we want

$$P(CD) \geq P^* \tag{1}$$

Two points need to made concerning the constant δ^*. First, we would like to correctly classify all those examinees having $\theta_i \geq \theta_o$ as well as those having $\theta_i < \theta_o$. Accordingly, the examiner will want to choose δ^* small so that there is negligible loss for misclassifying those examinees for whom $\theta_o - \delta^* < \theta_i < \theta_o + \delta^*$. However, we require $\delta^* > 0$ since otherwise it may be impossible to attain the desired goal of making a correct decision with probability at least P^*. Moreover, small values for δ^* will usually require a larger number of observations so that (1) is satisfied.

We observe that Tong (1969) has considered the problem of partitioning populations with respect to a control for the case $F(x; \theta_i)$ normal with mean θ_i and common variance σ^2. Our treatment of this problem considers a larger class of distribution functions which includes normal populations as a special case.

General Solution

We adopt the typical approach to ranking and selection problems, viz., we determine the vector $\underline{\theta}^o=(\theta_1^o,\ldots,\theta_k^o)$ for which the probability of a correct decision is a minimum. Once the vector $\underline{\theta}^o$ is determined, we may, in general, determine the required number of observations so that (1) is attained. In practice the vector $\underline{\theta}=(\theta_1,\ldots,\theta_k)$ will not equal $\underline{\theta}^o$ and so the

probability of a correct decision will usually be greater than P^*.

Since $\Psi(x_i)$ is an estimate of θ_i, we classify the examinees as follows: Assign the examinee to S_G if $\Psi(x_i) \geq \Psi(x_o)$ and to S_B otherwise.

Let $\theta_{k-g+1},\ldots,\theta_k$ be the g parameters greater than or equal $\theta_o + \delta^*$. Since the examinees for whom θ_i is in the indifference zone $(\theta_o-\delta^*, \theta_o+\delta^*)$ are classified correctly with probability one, it follows that to attain a lower bound on the probability of a correct decision, the remaining k-g parameters $\theta_1,\ldots,\theta_{k-g}$ must have a value less than or equal to $\theta_o-\delta^*$. Applying a theorem by Barr and Rizvi (1966), we have that the infimum of the probability of a correct decision occurs at $\theta_1=\theta_2=\cdots=\theta_{k-g}=\theta_o-\delta^*$, $\theta_{k-g+1}=\cdots=\theta_k=\theta_o+\delta^*$ and is given by

$$(2) \qquad \int_{-\infty}^{\infty} \prod_{i=1}^{k-g} F(x \mid \theta_i = \theta_o - \delta^*) \prod_{j=k-g+1}^{k} (1-F(x \mid \theta_j=\theta_o+\delta^*)\, dF(x;\theta_o)$$

To find the infimum of (2) over all possible values of θ_o it may be necessary to minimize (2) as a function of θ_o for all admissible values of g, i.e., g=0,1...,k. However, there are situations for which only one value of g need be considered. Suppose, for example, $dF(x; \theta_o)$ is symmetric about the origin. Since $F(x; \theta_i)$, i=1,...,k, is monotonically increasing in x, we may infer from a lemma given by Tong (1969, p. 1322) that the value of g which minimizes (2) is g=k/2 when k is even; g=(k-1)/2 or g=(k+1)/2 when k is odd.

Solution for the Binomial Error Model

We illustrate how the above analysis can be applied to the binomial error model (Lord, 1965; Lord and Novick, 1968). Under this model each examinee responds to n dichotomously scored test items and the probability that the i<u>th</u> examinee gets x correct is $P(x) = \binom{n}{x} \theta_i^x (1-\theta_i)^{n-x}$ where θ_i now represents the percent correct true score of the i<u>th</u> examinee. A natural and convenient choice for Ψ is $\Psi(x)=x$. In this case, $F(x; \theta_i) = \sum_{r=0}^{x} \binom{n}{r} \theta_i^r (1-\theta_i)^{n-r}$. We begin with the simplest case, namely, where the control is a single examinee and where $F(x; \theta_o)$ is also binomial. The results for this special case are of interest in their own right and they are also of use when we consider the more general situation where the control is a population of examinees.

We assume the examiner has specified the constants δ^* and P^* and that it is desired to choose n, the test length, so that the probability of a correct decision (PCD) is at least P^* for all those examinees for whom $\theta_i \leq \theta_o - \delta^*$ or $\theta_i \geq \theta_o + \delta^*$.

We first verify that the binomial distribution is stochastically increasing. As indicated by Sobel (1977)

$$\sum_{r=\alpha}^{n} \binom{n}{r} \theta^r (1-\theta)^{n-r} = I_\theta(\alpha, n-\alpha+1)$$

where I_θ is the usual incomplete beta function. Thus, for $\theta < \theta'$

$$\sum_{i=0}^{x} \binom{n}{i} \theta^i (1-\theta)^{n-i} = 1-I_\theta (x+1, n-x) \geq 1-I_{\theta'} (x+1, n-x) = \sum_{i=0}^{x} \binom{n}{i} (\theta')^i (1-\theta')^{n-i}$$

which is the desired result. It follows that the minimum probability of a

correct decision is given by (2) which for this special case becomes

$$(\hat{2}) \quad \sum_{x=0}^{n} [1-I_{\theta_o-\delta^*}(x, n-x+1)]^{k-g} [I_{\theta_o-\delta^*}(x, n-x+1)]^{g} \binom{n}{x} \theta_o^x (1-\theta_o)^{n-x}$$

Expression $(\hat{2})$ can be evaluated using subprogram BDTR in IBM's Scientific Subroutine Package (1971). By minimizing $(\hat{2})$ as a function of θ_o for $g=0(1)k$, we determine the minimum probability of a correct decision for a given n. By repeating this process for increasing values of n, we determine the shortest test length for which the desired probability guarantee is attained.

An Approximate Solution

We have found that the above procedure for determining the appropriate test length when comparing k examinees to a control examinee may be prohibitively expensive. We consider, therefore, an approximate solution which may be used in conjunction with a table given by Tong (1969).

The approximate solution we consider is based on the transformation

$$(3) \quad y = \tfrac{1}{2}[\sin^{-1}(\sqrt{x/N}) + \sin^{-1}(\sqrt{(x+1)/N+1})]$$

(Freeman and Tukey, 1950) where x is a binomial random variable with parameters N and θ. Temporarily we leave the integer K unspecified. The random variable y is approximately normally distributed with mean $\sin^{-1}(\sqrt{\theta})$ and variance $(4N+2)^{-1}$. Because the variance of a binomial distribution depends on the mean, inverse sine transformations such as (3) are used to stabilize the variance. Other arc-sine transformations exists but it is claimed by

Mosteller and Youtz (1961) as well as Mosteller and Tukey (1968) that (3) is currently the best angular transformation for variance stabilization for the binomial case. We also observe that the Freeman-Tukey transformation was suggested for use in a ranking and selection problem similar to the one under study (Gupta and Sobel, 1958). However, the accuracy of the resulting approximate solution was not investigated.

Before continuing, a point of clarification may be helpful. By using the Freeman-Tukey transformation (3), we may treat the observed scores of the k examinees as arising from k independent normal distributions having common variance $(4N+2)^{-1}$. Hence, as indicated below, we may use Tong's Table 1 (p. 1304) to determine M, the required number of observations so that (1) is satisfied. This means we take M sets of N observations, or n=MN binomial trials, so that the PCD is at least P^*.

Let $\delta_1^* = \sin^{-1}(\sqrt{\theta_o-\delta^*})-\sin^{-1}(\sqrt{\theta_o})$, $\delta_2^* = \sin^{-1}\sqrt{\theta_o+\delta^*}-\sin^{-1}(\sqrt{\theta_o})$

and set $a = (\delta_2^* - \delta_1^*)/2$. From Tong (1969, expression (1.10)) we see that the appropriate sample size is given approximately by choosing the smallest M such that

$$M \geq 2b^2/[a^2(4N+2)] \tag{4}$$

where the values of b are tabled by Tong for k=1(1)20, P^*=.50, .75, .90, .975, .99. Consequently, a conservative solution can be obtained by choosing the value of θ_o for which a is a minimum. Via differentiation this value is $\theta_o = \frac{1}{2}$. Thus, letting $c=(\sin^{-1}(\sqrt{.5+\delta^*})-\sin^{-1}(\sqrt{.5-\delta^*}))/2$, we choose the smallest M so that

(5) $$M \geq 2b^2/[c^2(4N+2)]$$

There remains the choice of N in (5). If we choose N too large, the resulting sample size may be larger than necessary. If N is too small, (3) may not be a sufficiently close approximation to a normal distribution. Initially we tried N=8 for determining the smallest M satisfying (5) with $\delta^* = .1$, $P^* = .75, .90, .95$, $k=1(1)5$. As a check on the accuracy of our approximate solution we used (2) to calculate the minimum probability of a correct decision with $n = N$. We found, however, that the resulting minimum probability of a correct decision did not always exceed P^*. Accordingly, we tried N=9, 10, 11, 12. The desired probability guarantee was attained in more but not all instances. Rather than increase N further, we tried setting $n=\max[n_1, n_2, n_3, n_4]$, where n_1, n_2, n_3, n_4 are the number of binomial trials corresponding to N=9(1)12. Our main justification for using this procedure is that we want to keep N small and because it seems to work fairly well in practice. Table 1 indicates the resulting values of n for $\delta^*=.1$, $P^* = .75, .90, .95$. Table 1 implies that if we administer a 72 item test to two examinees and a control examinee, there is at least a 75% chance of identifying correctly the examinees having true score greater than or equal to $\theta_o+.1$ (θ_o being the true score of the control examinee) or true score less than or equal to $\theta_o-.1$. Table 2 gives the actual minimum probability of a correct decision. The probabilities were computed using (3) and the values of n given by Table 1. Thus, the actual minimum probability of a correct decision based on a 72 item test for k=2 examinees is .774. It appears that the values of n using (5) tend to be too small as P^* gets close

to one. In practice it may be best to approximate the required test length using (5) and then adjust this value of n according to calculations made with (3).

Solution When the Control is a Population of Examinees

We generalize the above solution to the case where we have m number correct observed scores, say $z_1,\ldots z_m$ corresponding to m examinees randomly sampled from an appropriate control population. In this section we adopt the usual beta-binomial model. That is, we make the additional assumption that the probability function of true scores, say $g(\theta)$, for the control population is a beta distribution with parameters r and s, i.e.,

$$(6) \qquad g(\theta) = \frac{\Gamma(r+s)}{\Gamma(r)\,\Gamma(s)}\,\theta^{r-1}\,(1-\theta)^{s-1}$$

where Γ is the usual gamma function. (An unsubscripted greek θ refers to a true score in the control population and a subscripted greek θ refers to the true score of an examinee who is being compared to the control.) Not only is (6) motivated as the natural conjugate prior of the binomial distribution, there is also empirical evidence that (6) may provide a good approximation to the true score distribution (Keats and Lord, 1962). Assuming (6) holds, the marginal distribution of observed scores for the control population, say $h(z)$, is beta-binomial:

$$(7) \qquad h(z)=\binom{n}{z}\int_0^1 \frac{\Gamma(r+s)}{\Gamma(r)\,\Gamma(s)}\,\theta^{z+r-1}\,(1-\theta)^{n+s-z-1}$$

Performing the integration in (7), we have that

(8) $$h(z)=\binom{n}{z}\frac{\Gamma(r+s)\,\Gamma(r+z)\,\Gamma(n+s-z)}{\Gamma(r)\,\Gamma(s)\,\Gamma(n+r+s)}$$

It is desired to determine whether θ_i (the percent correct true score of the ith examinee) is above or below $\zeta = \frac{r}{r+s}$, the mean of the true score distribution $g(\theta)$. Since the mean of the beta-binomial distribution is $nr/(r+s)$, we make the decision $\theta_i \geq \frac{r}{r+s}$ when $x_i \geq \bar{z}$ where $\bar{z} = \Sigma\, z_i/m$ is the sample mean of the m control examinees. As before, the goal is to ensure that, with probability at least P^*, a correct decision is made for all examinees having $\theta_i \geq \zeta + \delta^*$ or $\theta_i \leq \zeta - \delta^*$.

A brief consideration of the case m=1 is both instructive and useful. Note that the derivation of (2) did not require $F(x;\, \theta_o)$ to be stochastically increasing in θ_o. Thus, we may derive an expression for the greatest lower bound on the PCD without verifying that (8) is stochastically increasing in ζ. From (2), this lower bound is

(9) $$\sum_{z=0}^{n} [1-I_{\zeta-\delta^*}(z,\, n-z+1)]^{k-g}\, [I_{\zeta+\delta^*}(z,\, n-z+1)]^{g}\, h(z)$$

where h(z) is given by (8).

Except for normal populations, it is difficult to make completely satisfactory progress beyond expression (9) when m is small. In fact it is not completely clear whether finite values for r and s exist which give a lower bound on (9). However, some progress can be made which may be useful in later studies. Exact results for normal populations are described in a later section. Also, more satisfactory results are possible when the number of sampled control examinees is large.

Earlier we noted that when h(z) is binomial, a lower bound on (9) occurs at or near ζ=.5. In a later section we show that when ζ is known, ζ=.5 again gives an approximate lower bound on the PCD. Both of these results suggest (albeit weakly) that a reasonable lower bound on (9) can be obtained by taking ζ=.5 which implies r=s. To lend some credence to our speculation, we evaluated (9) for many values of r and s. The parameters r and s where chosen primarily to include a variety of priors, namely, J-shaped, U-shaped, uniform, symmetric and skewed distributions. The resulting values of (9) are presented in Table 3 for k=5, g=3, n=10, 20, and (r, s) = (3, 3), (1, 1), (.5, .5), (9, 2), (2, 1), (3, 1). For all values of r and s used (not just those shown in Table 3) the two lowest PCD values occurred at (r, s) equal to (.5, .5) and (1, 1). Our results also suggest that even smaller PCD values will occur as r=s approaches zero.

We observe that (9) could be generalized to the case m > 1 by replacing h(z) with the probability function of $\bar{z}$. For the present, such an approach appears to offer little help in real life situations. When m is small, we merely compound the difficulties described for m=1. However, the approximate solution we are about to describe may be of some use.

Based on our empirical findings as well as the asymptotic results given earlier, it appears that the minimum PCD occurs at or near ζ=.5. These findings suggest that to obtain an approximate lower bound on the PCD, we assume ζ=.5, which implies r=s, and that we take the variance of the true score distribution to be as large as possible. For r=s, it is readily verified that the variance of the true score distribution approaches infinity as r approaches zero. Also, the variance of the prior is monotonically increasing as r(=s) gets small. But the beta distribution is defined

only for $r > 0$ and so the actual calculation of the minimum PCD is a bit awkward. However, by taking r reasonably close to zero, it should be possible to obtain a realistic lower bound on the PCD. One convenient choice is r=s=1, the uniform distribution. As already indicated, a smaller PCD can be obtained for smaller values of r and s. However, using r=s=1/2, for example, appears to have negligible effect on the PCD when m is of moderate size (see Table 4 below). Moreover, a U-shaped distribution (both r and s less than one) of true scores would seem to be highly unlikely in practice and so an assumption such as $r \geq 1$ and $s \geq 1$ appears to be both useful and realistic.

For the special case r=s=1, the marginal distribution of observed scores for a single examinee is the uniform distribution over the integers 0,1,...,n, namely,

$$f(z) = \frac{1}{n+1}, \quad z=0, 1, 2,\ldots,n.$$

The mean and variance are $\mu_z = n/2$ and $\sigma_z^2 = \frac{n(n+2)}{12}$. As before we let $\bar{z}$ denote the sample mean of observed test scores of the m control examinees. When m is not too small we may approximate the probability function of $\bar{z}$ with $w(\bar{z})$, say, via the central limit theorem. Let $[\bar{z}]$ denote the smallest integer greater than or equal to $\bar{z}$ and let $\{\bar{z}\}$ equal the greatest integer less than or equal to $\bar{z}$. From (2) and our discussion above, an approximate lower bound on the PCD is given by

$$(10) \qquad \sum_{\bar{z}} [1-I_{.5-\delta}(\{\bar{z}\}, n-\{\bar{z}\}+1)]^{k-g} [I_{.5+\delta}([\bar{z}], n-[\bar{z}]+1)]^{g} w(\bar{z}),$$

where $\bar{z}=0/m, 1/m, 2/m,\ldots,nm/m$. We calculated (10) for m=10, 50, 100, n=74, k=5, g=3. The results were .405, .635, .686. Using Monte Carlo techniques, we then estimated the actual PCD for a variety of priors. All estimates were based on 1000 iterations. The values in Table 3 were used to check the accuracy of the Monte Carlo experiments. The results are given in Table 4 for the priors used in creating Table 3. Observed that for m=50 and m=100 there is a negligible difference in the PCD for (r, s)=(.5, .5) and (r, s)=(1, 1). Even for m=10 the

difference is only about .05. All indications are that (10) gives a reasonable lower bound on the PCD.

Approximate solution when m is large

When m is large and n is not too small, we might assume that a reasonably accurate estimate of the mean of the true score distribution will be obtained. If we knew ζ, the problem of comparing examinees to a control reduces to the problem of comparing examinees to a standard for which a solution already exists (Wilcox, 1977). When ζ is known, the appropriate test length is given approximately by $n = \lambda^2 \zeta(1-\zeta)/(\delta^*)^2$ where λ is the standard normal table value corresponding to the $(P^*)^{1/k}$ percentile. One difficulty is that we do not know ζ prior to administering a test to the control examinees. One possibility is to assume ζ to be the value which would require the largest number of items on the test. By doing so, we have that (1) is satisfied no matter what the true value of ζ happens to be. In the asymptotic case it is readily verified that for the case ζ known, $\zeta = .5$ would require the largest number of test items so that (1) is satisfied.

Based on the above discussion, we propose the following approximate solution when m, the number of control examinees, is large. Determine the appropriate test length so that (1) is satisfied assuming $\zeta=.5$. After administering the test, compute $\bar{z}$, the mean of the observed test scores of the m control examinees. For each of the k examinees being compared to the control, assert that his percent correct true score, θ_i, is greater than or equal to the unknown true score of the control if his observed score is greater than or equal to $\bar{z}$. Otherwise we make the reverse decision. Further assert that a correct decision is made with probability at least P^*. Once ζ is estimated, the examiner might want to revise his estimate of the lower bound for the probability of a correct decision using the procedure given by Wilcox (1977).

To gain some insight as to how large m must be, suppose $\delta^* = .1$, k=5, and $P^* = .8$. If ζ is known to be .5, we have from Wilcox (1977) that the appropriate test length is n = 74 (which is why we used n=74 in the previous section). Referring to Table 4 we see that m=100 is not large enough to use this approximate solution since we set $P^* = .8$ and since the PCD is only .716 for r=s=1. We ran an additional Monte Carlo experiment with r=s=1, m=200. The estimated PCD was .776, a much more satisfactory result. For m=300 the PCD is .791. This illustrates that when using the approximate solution outlined above, m must be at least 300 and preferably larger. Of course the minimum m required is a function of n, the test length, as well as the distribution of true scores. In our example, m=300 appears to be conservative. However, if m < 300, it is better to evaluate (10) for a given value of n and all admissible values of g. If the value of (10) is $\geq P^*$ for each g, the test length n is satisfactory. Otherwise a larger value of n must be considered.

Solution for Normal Populations

Often it is possible to reduce or to approximate the measurements on examinees with a normal distribution and so the problem of comparing k normal populations to a control is of special interest. As indicated earlier, Tong (1969) has considered this problem for the case when the k normal populations and the control have a common variance. We indicate how this problem can be solved when the normal populations have unknown and unequal variances. Our procedure is based on a slight modification of a two-stage procedure by Dudewicz and Dalal (1975) where the goal is to select from k normal populations the one having the largest mean. As noted by Tong, it is impossible to get an exact solution using a one-stage procedure when the variances are unknown.

Let $h=h(k, P^*) > 0$ be the unique solution of the equation

$$(11) \qquad \int_{-\infty}^{\infty} (F_{n_o}(z+h))^{k-r} (1-F_{n_o}(z-h))^{r} f_{n_o}(z)dz = P^*$$

where $r=k/2$ if k is even and $r=(k-1)/2$ or $r=(k+1)/2$ if k is odd and where F_{n_o} and f_{n_o} are the distribution function and density function of a student's random variable with n_o-1 degrees of freedom.

Take an initial sample $X_{i1},\ldots,X_{in_o}$ from Π_i and define

$$\bar{X}_i(n_o) = \sum_{j=1}^{n_o} X_{ij}/n_o, \quad s_i^2 = \sum_{j=1}^{n_o} (X_{ij} - \bar{X}_i)^2/(n_o - 1)$$

$$n_i = \max \left\{n_o+1, \left[\left(\frac{s_i h}{\delta^*}\right)^2\right]\right\}$$

where as before $[y]$ denotes the smallest integer greater than or equal to y. Take n_i-n_o additional observations $X_{in_o+1},\ldots,X_{in_i}$ from Π_i and define $\tilde{X}_i=\sum_{j=1}^{n_i} a_{ij} X_{ij}$. The a_{ij}'s are chosen so that $a_i=\ldots=a_{in_i-1}=c_i$, $a_{in_i}=1-(n_i-1)c_i$, where

$$c_i = \frac{n_i-1 \pm \sqrt{(n_i-1)^2 - (n_i-1)n_i\,(1-(\delta^*/h)^2/s_i^2)}}{(n_i-1)n_i}$$

We make the decision $\theta_i \geq \theta_o$ if and only if $\tilde{X}_i \geq \tilde{X}_o$ where now θ_i $(i=0,1,\ldots,k)$ are the unknown means of the $k+1$ normal populations.

As before, let g denote the number of examinees having $\theta_i \geq \theta_o + \delta^*$. A straightforward modification of Dudewicz and Dalal's Theorem 4.1 shows that for a given g, the infimum of the probability of a correct decision is

$$(12) \qquad \int_{-\infty}^{\infty} (F_{n_o}(z+h))^{k-g} (1-F_{n_o}(z-h))^{g} f_{n_o}(z)dz.$$

Thus, to obtain the infimum over $\underline{\mu}=(\mu_1,\ldots,\mu_k)$ of the probability of a correct decision, it remains to determine the g which minimizes (12). From a lemma given by Tong (1969, p. 1322) which was mentioned earlier it follows that (12) is minimized by setting g=k/2 when k is even; g=(k-1)/2 or g=(k+1)/2 when k is odd. Since h was chosen to satisfy (11), the desired probability guarantee is attained.

Using Gauss-Legendre numerical quadrature, we have tabled the value of (11) for h=.1(.1)(5.1), k=1(1)24, n_o=1(1)15(5)30. These values may be used to attain the desired probability guarantee when partitioning k normal populations with respect to control. Our tables are quite lengthy though and so we did not include them here. Instead we have tabled the approximate values of h required for P^*=.75, .90, .95. The values of h given in Tables 5, 6 and 7 are conservative, i.e., they guarantee that the PCD is at least P^*. Moreover, for P^*=.75 they rarely over estimate the actual PCD by more than .02. For P^*=.95 the accuracy is better than .01 in every case and for P^*=.90 the accuracy is nearly always better than .01. Copies of our more extensive tables are available upon request.

One advantage of the two-stage procedure just described is that we are able to avoid the problems described earlier when m, the number of randomly sampled control examinees, is relatively small. If, as is customary, we assume a normal prior distribution having mean μ_o and variance τ^2, the marginal distribution of observed scores is known to be normal with mean μ_o and variance $\sigma^2+\tau^2$ where σ^2 is the variance of the conditional distribution of observed scores. Thus, we may use the two-stage procedure described above when σ^2 or τ^2 or both are unknown.

Some Concluding Remarks

It should be evident that the procedures described above can be extended to other strong true-score models which are of interest in certain testing situations. Consider the Poisson process models as described by Lord and Novick (1968). In this case

$$f(x|\theta) = e^{-\theta}\,\theta^{x}/x!, \tag{13}$$

the Poisson distribution, is used to described the observed responses of the examinees. It is readily verified that (13) is stochastically increasing in θ since

$$\sum_{i=0}^{x} f(i|\theta) = \frac{1}{x!}\int_{\theta}^{\infty} e^{-t}\,t^{n}\,dt$$

(Feller, 1968, p. 173) and so a lower bound on the probability of a correct decision is given by (2). When n, the number of trials, is not too small, we may also apply the results on normal distributions using the transformation

$$y = \sqrt{x + \frac{3}{8}} \tag{14}$$

where y is approximately normally distributed with mean $\sqrt{\theta}$ and variance 1/4.

Our final comment deals with the problem of choosing k in a particular investigation. A colleague suggested that determining test length in terms of k=1 examinee is most appropriate. If we set k=500, for example, an unrealistic test length is required for the minimum PCD to be reasonably close to one. Our objection to considering only k=1 (when there is really more than one examinee) is that the results may be misleading since for slightly larger values of k the PCD may drop dramatically. The problem seems to be

one of perspective. By considering various values of k, the investigator has a better understanding of the effect different test lengths have on the PCD. The statistician cannot tell the investigator which value of k to use just as he cannot tell the investigator what constitutes a small number. All the statistician can do is to report the minimum probability of making a correct decision in a given situation.

TABLE 1

Values of $n = \max [n_1, n_2, n_3, n_4]$ for $\delta^* = .1$

$P^* \backslash k$	1	2	3	4	5
.75	30	72	90	117	121
.90	84	135	160	180	198
.95	135	192	220	240	253

TABLE 2

Infimum of the probability of a correct decision using the value of n in Table 1

$P^* \backslash k$	1	2	3	4	5
.75	.742	.774	.757	.787	.759
.90	.889	.902	.897	.900	.900
.95	.943	.948	.946	.943	.940

TABLE 3

PCD values using (9) and m=1, k=5, g=3

(r, s)	N = 10	N = 20
(.5, .5)	.03810	.04784
(1, 1)	.05599	.07245
(2, 1)	.07834	.10169
(3, 1)	.10703	.13993
(3, 3)	.08588	.12011
(9, 2)	.13228	.26130

TABLE 4

Monte Carlo Estimates of PCD for
k=5, g=3, n=74

	M		
(r, s)	10	50	100
(.5, .5)	.384	.642	.698
(1, 1)	.431	.665	.716
(2, 1)	.532	.738	.797
(3, 1)	.660	.847	.869
(3, 3)	.562	.719	.776
(9, 2)	.821	.928	.935

TABLE 5

Approximate Values of h required for $P^* = .75$

n_0	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20	25	30
1	2.1	1.4	1.3	1.2	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.0	1.0	1.0
2	4.7	2.6	2.2	2.1	2.0	1.9	1.9	1.8	1.8	1.8	1.8	1.8	1.8	1.8	1.7	1.7	1.7
3	***	3.5	2.6	2.4	2.3	2.2	2.2	2.1	2.1	2.1	2.1	2.1	2.0	2.0	2.0	2.0	2.0
4	***	3.5	2.9	2.7	2.5	2.4	2.4	2.3	2.3	2.3	2.3	2.3	2.2	2.2	2.2	2.2	2.2
5	***	3.9	3.1	2.8	2.7	2.6	2.5	2.5	2.5	2.4	2.4	2.4	2.4	2.4	2.3	2.3	2.3
6	***	4.2	3.3	3.0	2.8	2.7	2.7	2.6	2.6	2.6	2.5	2.5	2.5	2.5	2.4	2.4	2.4
7	***	4.4	3.5	3.1	3.0	2.9	2.8	2.8	2.7	2.7	2.6	2.6	2.6	2.6	2.5	2.5	2.5
8	***	4.7	3.6	3.3	3.1	3.0	2.9	2.8	2.8	2.7	2.7	2.7	2.7	2.7	2.6	2.6	2.6
9	***	4.9	3.8	3.4	3.2	3.0	3.0	2.9	2.9	2.8	2.8	2.8	2.7	2.7	2.7	2.6	2.6
10	***	5.1	3.9	3.5	3.2	3.1	3.0	3.0	2.9	2.9	2.9	2.8	2.8	2.8	2.7	2.7	2.7
11	***	***	4.0	3.6	3.3	3.2	3.1	3.0	3.0	2.9	2.9	2.9	2.9	2.8	2.8	2.8	2.7
12	***	***	4.1	3.6	3.4	3.3	3.2	3.1	3.0	3.0	3.0	2.9	2.9	2.9	2.8	2.8	2.8
13	***	***	4.2	3.7	3.5	3.3	3.2	3.1	3.1	3.0	3.0	3.0	3.0	2.9	2.9	2.8	2.8
14	***	***	4.3	3.8	3.5	3.4	3.3	3.2	3.1	3.1	3.1	3.0	3.0	3.0	2.9	2.9	2.9
15	***	***	4.4	3.9	3.6	3.4	3.3	3.2	3.2	3.1	3.1	3.1	3.0	3.0	3.0	2.9	2.9
16	***	***	4.5	3.9	3.6	3.5	3.4	3.3	3.2	3.2	3.1	3.1	3.1	3.1	3.0	3.0	2.9
17	***	***	4.6	4.0	3.7	3.5	3.4	3.3	3.3	3.2	3.2	3.1	3.1	3.1	3.0	3.0	3.0
18	***	***	4.7	4.0	3.7	3.6	3.5	3.4	3.3	3.2	3.2	3.2	3.1	3.1	3.1	3.0	3.0
19	***	***	4.7	4.1	3.8	3.6	3.5	3.4	3.3	3.3	3.2	3.2	3.2	3.2	3.1	3.0	3.0
20	***	***	4.8	4.1	3.8	3.6	3.5	3.4	3.4	3.3	3.3	3.2	3.2	3.2	3.1	3.1	3.0

TABLE 6

Approximate Values of h for $P^* = .90$

k \ n_0	3	4	5	6	7	8	9	10	11	12	13	14	15	20	25	30
1	3.1	2.6	2.3	2.2	2.2	2.1	2.1	2.0	2.0	2.0	2.0	2.0	2.0	1.9	1.9	1.9
2	4.6	3.5	3.2	3.0	2.8	2.8	2.7	2.7	2.6	2.6	2.6	2.6	2.6	2.5	2.5	2.5
3	***	4.0	3.5	3.3	3.1	3.0	3.0	2.9	2.9	2.9	2.8	2.8	2.8	2.7	2.7	2.7
4	***	4.3	3.8	3.5	3.4	3.3	3.2	3.1	3.1	3.1	3.0	3.0	3.0	2.9	2.9	2.9
5	***	4.6	4.0	3.7	3.5	3.4	3.3	3.3	3.2	3.2	3.2	3.1	3.1	3.0	3.0	3.0
6	***	4.8	4.2	3.9	3.7	3.5	3.5	3.4	3.3	3.3	3.3	3.2	3.2	3.2	3.0	3.0
7	***	5.0	4.3	4.0	3.8	3.7	3.6	3.5	3.4	3.4	3.4	3.3	3.3	3.2	3.2	3.2
8	***	***	4.5	4.1	3.9	3.7	3.6	3.6	3.5	3.5	3.5	3.4	3.4	3.3	3.3	3.2
9	***	***	4.6	4.2	4.0	3.8	3.7	3.7	3.6	3.5	3.5	3.5	3.5	3.4	3.3	3.3
10	***	***	4.7	4.3	3.1	3.9	3.8	3.7	3.7	3.6	3.6	3.5	3.5	3.4	3.4	3.3
11	***	***	4.8	4.4	4.1	4.0	3.9	3.8	3.7	3.7	3.6	3.6	3.6	3.5	3.4	3.4
12	***	***	4.9	4.4	4.2	4.0	3.9	3.8	3.8	3.7	3.7	3.6	3.6	3.5	3.5	3.4
13	***	***	5.0	4.5	4.3	4.1	4.0	3.9	3.8	3.8	3.7	3.7	3.7	3.6	3.5	3.5
14	***	***	5.1	4.1	4.3	4.1	4.0	3.9	3.9	3.8	3.8	3.7	3.7	3.6	3.5	3.5
15	***	***	5.1	4.6	4.4	4.2	4.1	4.0	3.9	3.8	3.8	3.8	3.7	3.6	3.6	3.5
16	***	***	***	4.7	4.4	4.2	4.1	4.0	3.9	3.9	3.8	3.8	3.8	3.7	3.6	3.6
17	***	***	***	4.8	4.5	4.3	4.1	4.1	4.0	3.9	3.9	3.8	3.8	3.7	3.6	3.6
18	***	***	***	4.8	4.5	4.3	4.2	4.1	4.0	4.0	3.9	3.9	3.8	3.7	3.7	3.6
19	***	***	***	4.9	4.6	4.4	4.2	4.1	4.0	4.0	3.9	3.9	3.9	3.8	3.7	3.7
20	***	***	***	4.9	4.6	4.4	4.3	4.2	4.1	4.0	4.0	3.9	3.9	3.8	3.7	3.7

TABLE 7

Approximate Values of h for $P^* = .95$

k \ n_0	5	6	7	8	9	10	11	12	13	14	15	20	25	30
1	3.2	3.0	2.8	2.8	2.7	2.7	2.6	2.6	2.6	2.6	2.6	2.5	2.5	2.5
2	4.0	3.7	3.5	3.4	3.3	3.2	3.2	3.1	3.1	3.1	3.1	3.0	3.0	2.9
3	4.4	4.0	3.8	3.7	3.6	3.5	3.4	3.4	3.4	3.3	3.3	3.2	3.2	3.2
4	4.7	4.3	4.0	3.9	3.8	3.7	3.6	3.6	3.5	3.5	3.5	3.4	3.3	3.3
5	4.9	4.5	4.2	4.0	3.9	3.8	3.8	3.7	3.7	3.6	3.6	3.5	3.5	3.4
6	5.1	4.6	4.3	4.2	4.0	3.9	3.9	3.8	3.8	3.7	3.7	3.6	3.6	3.5
7	***	4.8	4.5	4.3	4.1	4.0	4.0	3.9	3.9	3.8	3.8	3.7	3.6	3.6
8	***	4.9	4.6	4.4	4.2	4.1	4.1	4.0	3.9	3.9	3.9	3.8	3.7	3.7
9	***	5.0	4.7	4.5	4.3	4.2	4.1	4.1	4.0	4.0	3.9	3.8	3.8	3.7
10	***	5.1	4.7	4.5	4.4	4.3	4.2	4.1	4.1	4.0	4.0	3.9	3.8	3.8
11	***	***	4.8	4.6	4.4	4.3	4.2	4.2	4.1	4.1	4.0	3.9	3.9	3.8
12	***	***	4.9	4.7	4.5	4.4	4.3	4.2	4.2	4.1	4.1	4.0	3.9	3.9
13	***	***	5.0	4.7	4.6	4.4	4.4	4.3	4.2	4.2	4.1	4.0	3.9	3.9
14	***	***	5.0	4.8	4.6	4.5	4.4	4.3	4.3	4.2	4.2	4.1	4.0	3.9
15	***	***	5.1	4.8	4.7	4.5	4.4	4.4	4.3	4.3	4.2	4.1	4.0	4.0
16	***	***	5.1	4.9	4.7	4.6	4.5	4.4	4.3	4.3	4.3	4.1	4.0	4.0
17	***	***	***	4.9	4.7	4.6	4.5	4.4	4.4	4.3	4.3	4.2	4.1	4.0
18	***	***	***	5.0	4.8	4.6	4.5	4.5	4.4	4.4	4.3	4.2	4.1	4.1
19	***	***	***	5.0	4.8	4.7	4.6	4.5	4.4	4.4	4.3	4.2	4.1	4.1
20	***	***	***	5.0	4.8	4.7	4.6	4.5	4.5	4.4	4.4	4.2	4.2	4.1

REFERENCES

Aitchison, J. and Aitken, C. G. G. Multivariate binary discrimination by the kernel method. Biometrika 1976, 63, 413-420.

Bahadur, R. R. A representation of the joint distribution of responses to n dichotomous items. In H. Solomon (Ed.) Studies in Item Analysis and Prediction. Stanford: Stanford University Press, 1961.

Barr, D. R. and Rizvi, M. H. An introduction to ranking and selection. Journal of the American Statistical Association, 1966, 61, 640-646.

Dudewicz, E. J. and Dalal, S. R. Allocation of observations in ranking and selection with unequal variances. Sankhya, 1975, Series B, 37, 28-78.

Feller, W. An Introduction to Probability Theory and Its Applications Vol. I, New York: John Wiley, 1968.

Freeman, M. F. and Tukey, J. W. Transformations related to the angular and the square root. The Annals of Mathematical Statistics, 1950, 21, 607-611.

Gupta, S. and Sobel, M. On selecting a subset which contains all populations better than a standard. Annals of Mathematical Statistics, 1958, 29, 235-244.

Huynh, [illegible] Statistical consideration of mastery scores. Psychometrika, 1[illegible], 41, 65-78.

IBM Application Program, System/360. Scientific subroutines package (360-CM-03X) Version III, programmer's manual. White Plains, New York: IBM Corporation Technical Publications Department, 1971.

Johns, M. V. An Empirical Bayes Approach to non-parametric two-way classification. In H. Solomon (Ed.) Studies in Item Analysis and Prediction, Stanford: Stanford University Press, 1961.

Keats, J. A. and Lord, F. M. A theoretical distribution for mental test scores. Psychometrika, 1962, 27, 59-72.

Lord, F. M. A strong true-score theory, with applications. Psychometrika, 1965, 30, 239-270.

Lord, F. M. and Novick, M. R. Statistical Theories of Mental Test Scores. Reading, Mass: Addison - Wesley, 1968.

Mosteller, F. and Tukey, J. W. Data analysis, including statistics. In G. Lindzey and E. Aronsen (Eds.). The Handbook of Social Psychology Reading, Mass: Addison - Wesley, 1968.

PARTITIONING NORMAL POPULATIONS WITH RESPECT TO A CONTROL

Rand Wilcox
Center for the Study of Evaluation
University of California at Los Angeles

August, 1977

ABSTRACT

Tong (1969) describes a two-stage procedure for partitioning k normal populations with respect to a control, the populations having a common but unknown variance. This problem can be solved for unknown and unequal variances with a modification of a procedure given by Dudewicz and Dalal (1975). This paper derives the appropriate probability equation to be solved so that a correct decision is made with probability at least P^*. Tables of the required h values are given for k=1(1)24, P^*=.75, .90, .95, .99 and n_0=2(1)15(5)30 where n_0 is the sample size in the first stage of the two-stage procedure.

1. INTRODUCTION

Using an indifference zone formulation, Tong (1969) describes ranking and selection procedures for partitioning k normal populations with respect to a control. The situation may be described as follows:
Let $\pi_0, \pi_1, \ldots, \pi_k$ be k+1 normal populations having common variance σ^2 and unknown means $\mu_0, \mu_1, \ldots, \mu_k$, respectively. Let $\delta^* > 0$ be an arbitrary but fixed constant and consider

$$\Omega_B = \{\pi_i : \mu_i < \mu_0 - \delta^*\}$$

$$\Omega_I = \{\pi_i : \mu_0 - \delta^* < \mu_i < \mu_0 + \delta^*\}$$

$$\Omega_G = \{\pi_i : \mu_i \geq \mu_0 + \delta^*\}.$$

After observations are taken the populations $\pi_1, \ldots, \pi_k$ are to be assigned to one of two disjoint sets, say S_B and S_G. A correct decision (CD) is made if every population in Ω_B is assigned to S_B and every population in Ω_G is assigned to S_G. No restriction is put on those populations in Ω_I; the populations in Ω_I are classified correctly with probability one. Let P^* be a preassigned constant such that $2^{-k} < P^* < 1$. The goal is to make a correct decision with probability at least P^*; the problem is to determine the minimum number of observations so that the desired probability guarantee is attained.

Given k normal populations having unknown and unequal variances, Dudewicz and Dalal (1975) derived a two-stage solution to the problem of selecting the normal population having the largest mean. Their solution

can be modified to yield a solution to the problem considered by Tong for the case of unequal variances. The purpose of this paper is to provide the appropriate probability equation to be solved as well as the tables needed to carry out the solution.

2. SOLUTION FOR UNKNOWN AND UNEQUAL VARIANCES

We now give a solution for the more general case in which the variances are unknown and not necessarily equal. To attain the desired goal we modify slightly the two-stage procedure of Dudewicz and Dalal (1975). Let $h=h(k, P^*) > 0$ be the unique solution of the equation

$$(1) \quad \int_{-\infty}^{\infty} F_\nu^{k-r}(z+h)(1-F_\nu(z-h))^r f_\nu(z)dz=P^*$$

where $r=k/2$ if k is even and $r=(k-1)/2$ or $r=(k+1)/2$ if k is odd and where F_ν and f_ν are the distribution function and density function of a Student's t distribution with $\nu=n_0-1$ degrees of freedom.

For the first stage, let $X_{i1},\ldots,X_{in_0}$ represent a sample of size n_0 from π_i and define

$$\bar{X}_i(n_0) = \sum_{j=1}^{n_0} X_{ij}/n_0, \quad s_i^2 = \sum_{j=1}^{n_0} (X_{ij} - \bar{X}_i)^2/(n_0-1)$$

$$n_i = \max\{n_0+1, [\left(\frac{s_i h}{\delta^*}\right)^2]\}$$

where $[y]$ denotes the smallest integer greater than or equal to y. For the second stage take n_i-n_0 additional observations $X_{in_0+1},\ldots,X_{in_i}$ from π_i and

define $\bar{X}_i = \sum_{j=1}^{n_i} a_{ij}$. The a_{ij}'s are chosen so that $a_{i1}=\ldots=a_{in_i-1}=c_i$, $a_{in_i}=1-(n_i-1)c_i$, where

$$c_i = \frac{n_i-1 \pm\sqrt{(n_i-1)^2 -(n_i-1)n_i[1-(\delta^*/h)^2/s_i^2]}}{(n_i-1)n_i}$$

Form the sets S_B and S_G as follows:

$$S_B = \{\pi_i: \bar{X}_i < \bar{X}_0\}$$

$$S_G = \{\pi_i: \bar{X}_i \geq \bar{X}_0\}$$

Theorem 1. Let $\pi_{k-g+1},\ldots,\pi_k$ be the g populations in Ω_G. We now show that for a given g, the infimum of the probability of a correct decision is

$$(2)\quad \int_{-\infty}^{\infty} F_\nu^{k-g}(z+h)(1-F_\nu(z-h))^g f_\nu(z)dz.$$

Proof. The proof of theorem 1 follows closely the proof of Dudewicz and Dalal's theorem 4.1. From the definition of a correct decision it follows that to obtain a lower bound on the probability of a correct decision, the set Ω_I must be empty and thus the remaining k-g populations, namely $\pi_1,\ldots,\pi_{k-g}$, are in Ω_B. For this special case we have that

$$(3)\quad P(CD) = P\{\bar{X}_i<\bar{X}_0,\ \bar{X}_j\geq\bar{X}_0\ (i=1,\ldots,k-g;\ j=k-g+1,\ldots,k)\}$$

$$= P\{\frac{\bar{X}_i-\mu_i}{\delta^*/h} < \frac{\bar{X}_0-\mu_0}{\delta^*/h} + \frac{\mu_0-\mu_i}{\delta^*/h},$$

$$\frac{\bar{X}_j-\mu_j}{\delta^*/h} \geq \frac{\bar{X}_0-\mu_0}{\delta^*/h} - \frac{\mu_j-\mu_0}{\delta^*/h}\ (i=1,\ldots,k-g;\ j=k-g+1,\ldots,k)\}$$

$$= \int_{-\infty}^{\infty} F_\nu^{k-g}(z+\frac{\mu_0-\mu_i}{\delta^*/h})(1-F_\nu(z-\frac{\mu_j-\mu_0}{\delta^*/h}))^g f_\nu(z)dz.$$

To establish the last equality in (3) we must show that the random variable

$$Y_i = \frac{\bar{X}_i - \mu_i}{\delta^*/h}$$

has a Student's-t distribution with $\nu = n_0 - 1$ degrees of freedom.

Let $w_i^2 = \sum_{j=1}^{n_i} a_{ij}^2$. Given s_i^2,

$$U_i = \frac{\bar{X}_i - \mu_i}{s_i w_i}$$

is normally distributed with mean zero and variance σ_i^2/s_i^2. From the choice of c_i it can be seen that

$$s_i^2 \sum_{j=1}^{n_i} a_{ij}^2 = (\delta^*/h)^2$$

As noted by Dudewicz and Dalal (1975) it can now be readily verified using standard techniques that Y_i has a Student's-t distribution with $n_0 - 1$ degrees of freedom.

To complete the proof we observe that for populations in Ω_B, $\mu_0 - \mu_i \geq \delta^*$, and for populations in Ω_G, $\mu_i - \mu_0 \geq \delta^*$, and so the infimum of (3) occurs at $\mu_0 - \mu_i = \delta^*$ $(i = 1, \ldots, k-g)$, and $\mu_j - \mu_0 = \delta^*$ $(j = k-g+1, \ldots, k)$.

To obtain the infimum over $\underline{\mu} = (\mu_1, \ldots, \mu_k)$ of the probability of a correct decision, it remains to determine the g which minimizes (2). Tong (1969, p. 1322) proves the following lemma. Let $s(z)$ and $S(z)$ be two real-valued functions. Let η be any positive constant and let λ be any real constant and consider

$$g(r) = \int_{-\infty}^{\infty} S^r(\eta z + \lambda)\, S^{k-r}(-\eta z + \lambda)\, f(z)\, dz$$

and its first difference

$$\Delta\beta(r) = \beta(r+1)-\beta(r)$$

If $S(z)$ is monotonically increasing, $S(z) \leq M$ for some $M > 0$, $s(z) \geq 0$, $\int_{-\infty}^{\infty} s(z)dz<\infty$ and if $s(z)=s(-z)$, then

$$\beta(r) = \beta(k-r),\ r=0,1,\ldots,k$$

and

$$\Delta\beta(r) \leq 0,\ r=0,1,\ldots,[(k-2)/2].$$

It is readily verified that the conditions of the lemma are met for f_ν and F_ν. It follows that the infimum of (2) occurs at $r=k/2$ if k is even; $g=(k+1)/2$ or $g=(k-1)/2$ if k is odd. Thus, the infimum of the probability of a correct decision is given by (1) as was to be shown.

3. CALCULATION OF THE TABLES

Calculation of our tables was accomplished in basically the same manner as the one used by Dudewicz and Dalal (1975). There are some differences though and so we describe our procedure in more detail.

We used 32-point Gauss-Legendre numerical quadrature (see, e.g., Krylov, 1962, pp. 107-111) to evaluate (1), the minimum probability of a correct decision. The necessary weights and evaluation points are given by Stroud and Secrest (1966). Since the integration in (1) is performed over an infinite range, we integrated (1) from $-d_1(n_0)$ to $d_2(n_0)$ where $d_1>0$ and $d_2>0$ were set so that the truncation error was at most 10^{-5}. This is possible by choosing d_1 and d_2 at least as large as the values used by Dudewicz and Dalal. (The actual values of d_1 and d_2 used were taken from another paper by Dudewicz, Ramberg and Chen, 1975.)

The tabled values of h are conservative in the sense that the left-hand side of (1) is at least P^*. Simultaneously, the left-hand side of (1) does not exceed P^* by more than 10^{-3}. The values of h were determined using a search routine based on the fact that (2) is a monotonically increasing function in h.

Due to the wide range of $(-d_1, d_2)$, we split the integral over $(-d_1, d_2)$ into several integrals, performed the Gauss-Legendre numerical quadrature on each, and added the results. We used 26 intervals for $n_0=2$, 22 intervals for $n_0=3$, 20 intervals for $n_0=3(1)14$, and 18 intervals for $n_0=20, 25, 30$.

Evaluation of (1) for the special case k=1, h=.1(.1)5.1, $n_0=2(1)15(5)30$ corresponds to the integral evaluated by Dudewicz and Dalal and serves as a check on our calculations. Our results agree with theirs except that on occasion, our value is .0001 larger than the value given by Dudewicz and Dalal. The reason for the discrepancy is that they used 64 points in their numerical quadrature rather than the 32 points used here. However, their results are not necessarily more accurate than ours since we integrated over a larger number of intervals. Moreover, Dudewicz and Dalal report that their results are correct up to $\pm$ .0001 and so this is a negligible discrepancy.

Numerical Illustration

To illustrate the solution, consider four populations Π_0, Π_1, Π_2 and Π_3 where, unknown to the investigator, the corresponding random variables are $X_0 \sim N(0,1)$, $X_1 \sim N(-2, 1/4)$, $X_2 \sim N(2,4)$ and $X_3 \sim N(1.5, 4)$. Let $\delta^*=1, P^*=.9$ and $n_0 = 10$. Thus, a correct decision assigns Π_2 and Π_3 to S_G and Π_1 to S_B. From the table headed $P^* = .9$ we see from the row $k=3$ and the column $n_0=10$ that the required value of h is 2.88.

Table 1 gives artificially generated data for the first stage of the two-stage solution. Summary statistics are listed at the bottom of the table. From section 2 we see that $n_0=11$, $n_1=11$, $n_2=38$, and $n_3=13$. Thus, for the second stage, we need one additional observation from both Π_0 and Π_1, 28 observations from Π_3 and 3 observations from Π_4. To compute $\bar{X}_i$, we may use $c_0=.1207$, $c_1=.1519$, $c_2=.0267$, and $c_3=.0327$. Using the artificially generated data in Table 2 for the second stage, we have $\bar{X}_0=.3263$, $\bar{X}_1=-1.563$, $\bar{X}_2=2.606$, and $\bar{X}_3=2.759$.

TABLE 1

Artificially generated data for the first stage of the two-stage procedure

i	Π_0	Π_1	Π_2	Π_3
1	-1.10	-.11	6.74	1.36
2	.24	-.88	2.68	2.50
3	.59	-.78	7.40	4.46
4	1.22	-1.20	4.92	4.84
5	.53	-.97	1.62	2.54
6	-.11	-1.45	5.60	1.62
7	1.13	-1.54	5.05	2.80
8	-.25	-1.87	2.48	4.42
9	-1.08	-1.16	3.74	2.30
10	.35	-1.44	1.22	2.92
Σx	1.52	-11.4	41.46	29.76
s^2	.6408	.2411	4.544	1.458

TABLE 2

Artificially generated data for the second stage of the two-stage procedure

i	Π_0	Π_1	Π_2	Π_3
11	-.62	-.325	3.84	.96
12			.03	2.10
13			4.52	5.86
14			2.42	
15			4.42	
16			.56	
17			.84	
18			1.10	
19			7.58	
20			2.74	
21			2.93	
22			6.93	
23			2.40	
24			3.35	
25			-0.92	
26			2.05	
27			.74	
28			1.52	

Values of h for $P^* = .75$

k \ n_0	2	3	4	5	6	7	8	9	10
1	2.01	1.37	1.21	1.14	1.10	1.072	1.054	1.041	1.031
2	4.69	2.58	2.173	2.01	1.92	1.862	1.824	1.796	1.775
3	6.05	3.38	2.55	2.341	2.23	2.16	2.112	2.08	2.052
4	7.37	3.497	2.86	2.61	2.476	2.394	2.34	2.299	2.27
5	8.55	3.83	3.081	2.795	2.645	2.553	2.492	2.45	2.414
6	9.73	4.13	3.28	2.961	2.795	2.694	2.626	2.58	2.54
7	10.9	4.39	3.45	3.094	2.914	2.804	2.73	2.68	2.64
8	12.02	4.63	3.60	3.22	3.022	2.905	2.826	2.77	2.73
9	13.2	4.85	3.73	3.322	3.114	2.99	2.905	2.846	2.801
10	14.3	5.07	3.86	3.421	3.201	3.07	2.98	2.92	2.87
11	15.41	5.26	3.97	3.51	3.28	3.14	3.044	2.98	2.93
12	16.53	5.45	4.08	3.592	3.35	3.202	3.105	3.04	2.986
13	17.7	5.63	4.18	3.67	3.412	3.26	3.16	3.09	3.035
14	18.8	5.81	4.28	3.74	3.474	3.316	3.212	3.14	3.083
15	19.9	5.98	4.363	3.81	3.53	3.366	3.26	3.182	3.125
16	21.0	6.14	4.45	3.871	3.584	3.415	3.303	3.225	3.166
17	22.11	6.29	4.531	3.93	3.634	3.46	3.345	3.264	3.204
18	23.3	6.45	4.61	3.99	3.683	3.503	3.385	3.301	3.24
19	24.4	6.60	4.69	4.042	3.73	3.543	3.422	3.336	3.273
20	25.5	6.75	4.76	4.10	3.772	3.582	3.46	3.37	3.305
21	26.6	6.88	4.83	4.15	3.813	3.62	3.491	3.402	3.335
22	27.7	7.02	4.896	4.194	3.853	3.654	3.524	3.432	3.364
23	28.8	7.15	4.961	4.24	3.891	3.69	3.554	3.461	3.392
24	29.9	7.28	5.03	4.29	3.93	3.72	3.584	3.49	3.42

Values of h for $P^* = .75$ (continued)

k \ n_0	11	12	13	14	15	20	25	30
1	1.023	1.02	1.011	1.01	1.003	.99	.982	.98
2	1.76	1.745	1.734	1.725	1.72	1.691	1.676	1.666
3	2.032	2.015	2.002	1.991	1.981	1.95	1.932	1.92
4	2.246	2.227	2.211	2.198	2.19	2.151	2.13	2.117
5	2.39	2.366	2.35	2.335	2.323	2.282	2.25	2.245
6	2.512	2.49	2.47	2.455	2.441	2.397	2.372	2.357
7	2.61	2.583	2.563	2.546	2.532	2.485	2.46	2.441
8	2.695	2.67	2.65	2.63	2.614	2.564	2.536	2.52
9	2.767	2.74	2.716	2.70	2.682	2.63	2.60	2.58
10	2.834	2.805	2.781	2.761	2.745	2.69	2.66	2.64
11	2.891	2.861	2.836	2.816	2.80	2.741	2.71	2.69
12	2.946	2.914	2.89	2.87	2.85	2.79	2.756	2.735
13	2.994	2.961	2.935	2.913	2.894	2.832	2.797	2.775
14	3.04	3.006	2.98	2.956	2.936	2.873	2.837	2.814
15	3.081	3.046	3.02	2.994	2.975	2.91	2.872	2.85
16	3.121	3.085	3.056	3.032	3.011	2.944	2.906	2.882
17	3.16	3.12	3.09	3.065	3.044	2.975	2.936	2.912
18	3.192	3.154	3.123	3.10	3.076	3.005	2.966	2.941
19	3.224	3.185	3.154	3.13	3.106	3.033	2.993	2.967
20	3.255	3.216	3.183	3.157	3.134	3.06	3.02	2.993
21	3.284	3.244	3.211	3.183	3.161	3.085	3.043	3.016
22	3.312	3.271	3.24	3.21	3.185	3.11	3.066	3.04
23	3.34	3.296	3.262	3.234	3.21	3.132	3.09	3.06
24	3.364	3.321	3.286	3.26	3.233	3.154	3.11	3.08

Values of h for $P^* = .90$

k \ n_0	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20	25	30
1	6.2	3.09	2.51	2.292	2.18	2.11	2.06	2.03	2.00	1.98	1.97	1.95	1.94	1.93	1.897	1.88	1.87
2	9.1	4.6	3.49	3.11	2.91	2.79	2.711	2.66	2.62	2.581	2.56	2.54	2.52	2.502	2.452	2.43	2.41
3	15.9	5.25	3.93	3.46	3.23	3.09	2.993	2.93	2.88	2.841	2.811	2.79	2.77	2.75	2.692	2.66	2.64
4	22.3	5.84	4.28	3.75	3.48	3.32	3.21	3.14	3.081	3.04	3.01	2.98	2.96	2.94	2.872	2.84	2.814
5	25.4	6.32	4.55	3.95	3.66	3.48	3.37	3.29	3.23	3.18	3.14	3.11	3.09	3.07	2.996	2.96	2.933
6	28.5	6.76	4.79	4.13	3.81	3.62	3.496	3.41	3.35	3.295	3.26	3.23	3.197	3.18	3.102	3.061	3.035
7	31.5	7.16	4.99	4.28	3.94	3.74	3.601	3.51	3.441	3.39	3.35	3.313	3.29	3.262	3.19	3.141	3.114
8	34.6	7.53	5.19	4.42	4.05	3.84	3.70	3.60	3.53	3.473	3.43	3.40	3.37	3.34	3.26	3.213	3.19
9	35.0	7.9	5.35	4.54	4.15	3.93	3.78	3.68	3.601	3.55	3.50	3.461	3.43	3.41	3.32	3.273	3.244
10	37.6	8.21	5.51	4.65	4.25	4.01	3.86	3.75	3.67	3.61	3.562	3.523	3.491	3.47	3.38	3.33	3.30
11	41.0	8.52	5.66	4.75	4.33	4.08	3.92	3.81	3.73	3.67	3.62	3.58	3.55	3.52	3.43	3.38	3.35
12	44.0	8.82	5.79	4.85	4.40	4.15	3.99	3.87	3.79	3.72	3.67	3.63	3.60	3.57	3.473	3.421	3.39
13	47.0	9.1	5.92	4.94	4.48	4.21	4.04	3.92	3.84	3.77	3.72	3.673	3.64	3.61	3.513	3.461	3.43
14	50.0	9.4	6.05	5.02	4.54	4.27	4.091	3.97	3.88	3.813	3.76	3.72	3.68	3.65	3.552	3.50	3.464
15	53.0	9.65	6.16	5.10	4.60	4.32	4.14	4.02	3.93	3.86	3.80	3.76	3.72	3.69	3.59	3.532	3.50
16	56.0	9.91	6.27	5.17	4.66	4.37	4.19	4.06	3.97	3.893	3.84	3.791	3.76	3.722	3.62	3.564	3.53
17	59.0	10.2	6.38	5.24	4.72	4.42	4.23	4.10	4.002	3.93	3.871	3.83	3.79	3.76	3.65	3.593	3.56
18	62.0	10.4	6.48	5.31	4.77	4.47	4.27	4.14	4.04	3.97	3.91	3.86	3.82	3.79	3.68	3.621	3.584
19	65.0	10.64	6.58	5.37	4.82	4.51	4.31	4.171	4.071	4.00	3.94	3.89	3.85	3.82	3.71	3.65	3.61
20	68.0	10.9	6.68	5.44	4.87	4.55	4.35	4.21	4.11	4.03	3.97	3.92	3.88	3.842	3.732	3.672	3.633
21	71.0	11.1	6.77	5.49	4.92	4.59	4.38	4.24	4.14	4.06	4.00	3.95	3.902	3.87	3.76	3.70	3.66
22	74.0	11.32	6.86	5.55	4.96	4.63	4.42	4.27	4.17	4.09	4.02	3.97	3.93	3.893	3.78	3.72	3.68
23	77.0	11.53	6.94	5.61	5.00	4.66	4.45	4.30	4.191	4.11	4.05	4.00	3.952	3.92	3.801	3.74	3.70
24	80.1	11.8	7.03	5.66	5.04	4.70	4.48	4.33	4.22	4.14	4.07	4.02	3.98	3.94	3.822	3.76	3.72

Values of h for $P^* = .95$

k \ n_0	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20	25	30
1	12.7	4.59	3.497	3.11	2.97	2.793	2.72	2.66	2.62	2.59	2.56	2.54	2.52	2.51	2.46	2.43	2.41
2	25.4	6.54	4.59	3.94	3.63	3.45	3.33	3.25	3.19	3.14	3.10	3.07	3.04	3.018	2.95	2.91	2.89
3	32.0	7.5	5.09	4.33	3.96	3.75	3.61	3.52	3.45	3.39	3.35	3.31	3.29	3.26	3.18	3.14	3.11
4	39.0	8.3	5.496	4.63	4.22	3.98	3.83	3.72	3.64	3.58	3.54	3.493	3.47	3.44	3.35	3.298	3.27
5	45.0	8.9	5.82	4.86	4.41	4.15	3.98	3.87	3.78	3.72	3.67	3.63	3.59	3.56	3.47	3.42	3.39
6	52.0	9.51	6.10	5.06	4.57	4.291	4.12	3.99	3.90	3.84	3.78	3.74	3.697	3.67	3.57	3.52	3.48
7	58.0	10.1	6.35	5.22	4.71	4.41	4.23	4.10	3.997	3.93	3.87	3.83	3.79	3.76	3.65	3.59	3.56
8	64.0	10.6	6.57	5.38	4.83	4.52	4.32	4.19	4.09	4.01	3.95	3.90	3.86	3.83	3.72	3.66	3.63
9	70.0	11.1	6.78	5.51	4.94	4.61	4.41	4.26	4.16	4.08	4.02	3.97	3.93	3.90	3.78	3.72	3.68
10	77.0	11.6	6.97	5.64	5.03	4.70	4.48	4.34	4.23	4.15	4.08	4.03	3.99	3.95	3.84	3.78	3.74
11	83.0	12.0	7.14	5.75	5.12	4.77	4.55	4.40	4.29	4.20	4.14	4.09	4.04	4.01	3.89	3.82	3.78
12	89.0	12.4	7.31	5.86	5.21	4.84	4.62	4.46	4.34	4.26	4.19	4.14	4.09	4.05	3.93	3.86	3.82
13	95.0	12.8	7.47	5.96	5.28	4.91	4.67	4.51	4.39	4.31	4.24	4.19	4.13	4.10	3.97	3.90	3.86
14	101.0	13.2	7.62	6.05	5.35	4.97	4.73	4.56	4.44	4.35	4.28	4.22	4.18	4.14	4.01	3.94	3.90
15	108.0	13.6	7.76	6.14	5.42	5.03	4.78	4.61	4.49	4.39	4.32	4.26	4.21	4.17	4.04	3.97	3.93
16	114.0	14.0	7.895	6.22	5.49	5.08	4.83	4.65	4.53	4.43	4.36	4.30	4.25	4.21	4.08	4.00	3.96
17	120.0	14.3	8.03	6.30	5.55	5.13	4.87	4.69	4.56	4.47	4.39	4.33	4.28	4.24	4.11	4.03	3.99
18	126.0	14.7	8.16	6.38	5.60	5.18	4.91	4.73	4.60	4.50	4.43	4.36	4.31	4.27	4.13	4.06	4.01
19	133.0	15.0	8.28	6.45	5.66	5.22	4.95	4.77	4.64	4.53	4.46	4.39	4.34	4.30	4.16	4.08	4.04
20	139.0	15.4	8.39	6.52	5.71	5.27	4.99	4.80	4.67	4.57	4.49	4.42	4.37	4.33	4.19	4.11	4.06
21	145.0	15.7	8.51	6.59	5.76	5.31	5.03	4.84	4.70	4.60	4.51	4.45	4.40	4.35	4.21	4.13	4.08
22	151.0	16.0	8.62	6.66	5.81	5.35	5.06	4.87	4.73	4.62	4.54	4.48	4.42	4.38	4.23	4.15	4.10
23	157.0	16.3	8.73	6.72	5.86	5.39	5.10	4.90	4.76	4.65	4.57	4.50	4.45	4.40	4.25	4.17	4.12
24	164.0	16.6	8.83	6.78	5.90	5.42	5.13	4.93	4.78	4.68	4.59	4.52	4.47	4.42	4.27	4.19	4.14

Values of h for $P^* = .99$

k \ n_0	3	4	5	6	7	8	9	10	11	12	13	14	15	20	25	30
1	10.30	6.40	5.20	4.70	4.31	4.12	4.00	3.92	3.82	3.80	3.72	3.70	3.70	3.60	3.50	3.50
2	14.50	8.00	6.20	5.40	5.00	4.72	4.60	4.42	4.32	4.30	4.20	4.13	4.10	4.00	3.90	3.85
3	16.30	8.70	6.70	5.80	5.30	5.01	4.81	4.70	4.60	4.50	4.42	4.40	4.32	4.20	4.10	4.05
4	17.90	9.30	7.10	6.10	5.60	5.30	5.02	4.90	4.80	4.70	4.60	4.53	4.50	4.33	4.24	4.20
5	19.40	9.80	7.40	6.30	5.80	5.40	5.20	5.01	4.90	4.80	4.72	4.70	4.61	4.44	4.40	4.30
6	20.70	10.30	7.60	6.50	5.90	5.60	5.30	5.20	5.00	4.90	4.83	4.80	4.71	4.53	4.44	4.40
7	21.90	10.60	7.80	6.70	6.10	5.70	5.42	5.30	5.10	5.00	4.92	4.90	4.80	4.61	4.51	4.50
8	23.0	11.00	8.10	6.80	6.20	5.80	5.51	5.32	5.20	5.10	5.00	4.92	4.90	4.70	4.60	4.52
9	25.0	11.30	8.20	7.00	6.30	5.90	5.60	5.40	5.30	5.20	5.10	5.00	4.93	4.74	4.63	4.60
10	26.0	11.70	8.40	7.10	6.40	6.00	5.70	5.50	5.32	5.21	5.12	5.10	5.00	4.80	4.70	4.62
11	27.0	11.90	8.60	7.20	6.50	6.10	5.80	5.60	5.40	5.30	5.20	5.10	5.03	4.83	4.73	4.70
12	28.0	12.20	8.70	7.30	6.60	6.10	5.80	5.60	5.43	5.31	5.22	5.20	5.10	4.90	4.80	4.70
13	28.0	12.50	8.90	7.40	6.60	6.20	5.90	5.70	5.50	5.40	5.30	5.20	5.13	4.92	4.90	4.73
14	29.0	12.70	9.00	7.50	6.70	6.21	5.90	5.70	5.63	5.40	5.31	5.23	5.20	5.00	4.84	4.80
15	30.0	13.00	9.10	7.60	6.80	6.30	6.00	5.80	5.60	5.50	5.40	5.30	5.20	5.00	4.90	4.80
16	31.0	13.20	9.20	7.70	6.80	6.40	6.00	5.80	5.61	5.50	5.40	5.30	5.23	5.01	4.90	4.83
17	32.0	13.40	9.30	7.70	6.90	6.40	6.10	5.82	5.70	5.52	5.42	5.33	5.30	5.04	4.93	4.85
18	33.0	13.60	9.40	7.80	7.00	6.50	6.10	5.90	5.70	5.60	5.50	5.40	5.30	5.10	5.00	4.90
19	33.0	13.90	9.60	7.90	7.00	6.50	6.20	5.90	5.72	5.60	5.50	5.40	5.32	5.10	5.00	4.90
20	34.0	14.10	9.70	7.90	7.10	6.50	6.20	5.93	5.80	5.61	5.51	5.42	5.40	5.12	5.00	4.92
21	35.0	14.20	9.80	8.00	7.10	6.60	6.20	6.00	5.80	5.70	5.53	5.50	5.40	5.14	5.02	4.94
22	36.0	14.40	9.80	8.10	7.20	6.60	6.30	6.00	5.81	5.70	5.60	5.50	5.40	5.20	5.04	5.00
23	37.0	14.60	9.90	8.10	7.20	6.70	6.30	6.02	5.90	5.70	5.60	5.50	5.42	5.20	5.10	5.00
24	37.0	14.80	10.00	8.20	7.30	6.70	6.30	6.10	5.90	5.72	5.61	5.52	5.50	5.21	5.10	5.00

REFERENCES

Dudewicz, E. J. and Dalal, S. R. Allocation of observations in ranking and selection with unequal variances. Sankhya, 1975, Series B, 37, 28-78.

Dudewicz, E. J., and Ramberg, J. S. and Chen, H. J. New tables for multiple comparisons with a control (unknown variances). Biometrische Zeitschrift, 1975, 17, 13-26.

Krylov, V. I. Approximate Calculation of Integrals. New York: Macmillan, 1962.

Stroud, A. H., and Secrest, D. Gaussian Quadrature Formulas. New Jersey: Prentice-Hall, 1966.

Tong, Y. L. On partitioning a set of normal populations by their locations with respect to a control. Anuals of Mathematical Statistics, 1969, 40, 1300-1324.